Analyse longitudinale de la qualité de vie relative à la santé en cancérologie

Amelie Anota

UNIVERSITE DE FRANCHE-COMTE

ECOLE DOCTORALE « ENVIRONNEMENTS-SANTE »

Année 2014

THESE

Pour obtenir le grade de

Docteur de l'Université de Franche-Comté

Discipline : Sciences de la Vie et de la Santé

Spécialité : Biostatistiques

Présentée et soutenue publiquement

Le 22 octobre 2014

Par

Amélie ANOTA

Analyse longitudinale de la qualité de vie relative à la santé en cancérologie

Directeur de thèse : Pr Franck Bonnetain

JURY

Dr Caroline Bascoul-Mollevi, PhD, Intitut du Cancer de Montpellier	(co-encadrante de thèse)
Pr Franck Bonnetain, PU-PH, Université de Franche-Comté	(directeur de thèse)
Dr Fabio Efficace, PhD, GIMEMA, Italie	(examinateur)
Pr Florence Joly, PU-PH, Université de Caen	(rapporteur)
Pr Célestin Kokonendji, PU, Université de Franche-Comté	(examinateur)
Pr Christian Lavergne, PU, Université de Montpellier 3	(examinateur)
Pr Mounir Mesbah, PU, Université de Paris VI	(rapporteur)
Pr Florence Tubach, PU-PH, Université Paris VII	(examinateur)

Remerciements

Je tiens tout d'abord à remercier mon Directeur de thèse, Monsieur Franck Bonnetain, Professeur à l'Université de Franche-Comté, pour m'avoir donné l'opportunité de mener ce travail de recherche, pour sa confiance, ses encouragements et ses observations critiques, dans l'élaboration de ce travail. Je le remercie de m'avoir fait découvrir et partager ses connaissances sur la thématique de la qualité de vie, durant mon stage de master, et de m'avoir permis de poursuivre mes recherches en thèse, tout en m'accordant une très grande liberté dans la réalisation de mes travaux.

Je remercie ma co-encadrante de thèse, Madame Caroline Bascoul-Mollevi, Chercheur à l'Institut du Cancer de Montpellier, pour ses conseils avisés, sa rigueur scientifique, son enthousiasme et son soutien, tout au long de ce travail.

Je remercie Madame Florence Joly, Professeur à l'Université de Caen Basse-Normandie et Monsieur Mounir Mesbah, Professeur à l'Université Pierre et Marie Curie pour avoir accepté de lire, d'analyser et de juger ce travail de thèse et d'en être rapporteurs auprès de l'Université. Leurs observations critiques m'ont été très utiles.

J'adresse ma gratitude à Madame Florence Tubach, Professeur à l'Université Paris Diderot, à Monsieur Célestin Kokonendji, Professeur à l'Université de Franche-Comté, à Monsieur Christian Lavergne, Professeur à l'Université de Montpellier ainsi qu'à Monsieur Fabio Efficace, Chercheur à GIMEMA, qui me font l'honneur d'évaluer ce travail en tant que membres de mon jury de thèse.

Mes remerciements s'adressent également aux membres de la Plateforme Qualité de Vie et Cancer avec lesquels j'ai le plaisir de travailler quotidiennement. Plus particulièrement, je remercie Sophie Parnalland, pour ses conseils, son soutien et pour avoir consacré du temps à la relecture de ce travail.

Je tiens également à remercier mes collègues de l'Unité de Méthodologie et de Qualité de Vie en Cancérologie du CHU de Besançon. Mes remerciements s'adressent en particulier à Astrid, Marie et Morgane, pour leur soutien, leurs conseils ainsi que pour tous les bons moments que nous avons passés ensemble.

Enfin, je tiens à remercier ma famille et mes amis pour m'avoir soutenu durant toutes ces années.

Cette thèse a été en partie financée par l'IRESP (Institut de Recherche en Santé Publique).

Table des matières

Liste des tableaux

Table des figures

Abréviations

ACP	Analyse en Composantes Principales
AFCM	Analyse Factorielle des Correspondances Multiples
AIC	Akaike Information Criteria
ANOVA	Analyse de variance
AR(1)	First-order autoregressive
ARH(1)	Heterogeneous first-order autoregressive
ASCO	American Society of Clinical Oncology
AUC	Area Under the Curve
CCI	Coefficient de Corrélation Intra-classe
CONSORT	Consolidated Standards of Reporting Trials
CSH	Heterogeneous Compound Symmetry
CTT	Classical Test Theory
DIF	Differential Item Functioning
DMCI	Différence minimale cliniquement importante
EORTC	European Organization of Research and Treatment of Cancer
EPIC	Expanded Prostate Index Cancer
FACT	Functional Assessment of Cancer Therapy
FDA	Food and Drug Administration
FLIC	Functional Living Index Cancer
GEE	Generalized Estimating Equation
GPCM	Generalized Partial Credit Model
GRM	Graded Response Model
ICC	Item Characteristic Curve
IRT	Item Response Theory
LASA	Linear Analogue Self-Assessment
LLRA	Linear Logistic Model with Relaxed Assumptions
LLTM	Linear Logistic Test Model
LOCF	Last Observation Carried Forward
LPCM	Longitudinal Partial Credit Model
MAR	Missing At Random
MCAR	Missing Completely At Random
MNAR	Missing Not At Random
NCI-CTC AE	National Cancer Institute Common Terminology Criteria for Adverse Events
OMS	Organisation Mondiale de la Santé
PCM	Partial Credit Model
PROs	Patient-Reported Outcomes
QdV	Qualité de Vie relative à la santé
QALYs	Quality-Adjusted Life Years
Q-TWIST	Quality-adjusted Time Without Symptoms Toxicity
RMST	Restricted Mean Survival Time
RSM	Rating Scale Model

SEIQoL	Schedule for the Evaluation of Individual Quality of Life
SEM	Structural Equation Modeling
TJD	Temps jusqu'à détérioration
TJDD	Temps jusqu'à détérioration définitive
TOI	Trial Outcome Index
UN	Unstructured
WHOQOL	World Health Organization Quality Of Life

Productions scientifiques

Publications

- Kepka S, Baumann C, Anota A, Buron G, Spitz E, Auquier P, Guillemin F, Mercier M. The relationship between traits optimism and anxiety and health-related quality of life in patients hospitalized for chronic diseases: data from the SATISQOL study. Health Qual Life Outcomes. 2013 Aug 5;11:134. doi: 10.1186/1477-7525-11-134

- Anota A, Hamidou Z, Paget-Bailly S, Chibaudel B, Bascoul-Mollevi C, Auquier P, Westeel V, Fiteni F, Borg C, Bonnetain F. Time to Health-related Quality of Life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization? Qual Life Res. 2013 Nov 26

- Nguyen TV, Anota A, Brédart A, Monnier A, Bosset JF, Mercier M. A longitudinal analysis of patient satisfaction with care and quality of life in ambulatory oncology based on the OUT-PATSAT35 questionnaire. BMC Cancer. 2014 Jan 25;14:42. doi: 10.1186/1471-2407-14-42.

- Boyer L, Baumstarck K, Michel P, Boucekine M, Anota A, Bonnetain F, Coste J, Falissard B, Guilleux A, Hardouin JB, Loundou A, Mercier M, Mesbah M, Rouquette A, Sebille V, Verdam MG, Ghattas B, Guillemin F, Auquier P. Statistical challenges of quality of life and cancer: new avenues for future research. Expert Rev Pharmacoecon Outcomes Res 2014 Feb;14(1):19-22

- Panouillères M, Anota A, Nguyen TV, Brédart A, Bosset JF, Monnier A, Mercier M, Hardouin JB. Evaluation properties of the French version of the OUT-PATSAT35 satisfaction with care questionnaire according to classical and item response theory analyses. Qual Life Res. 2014 Sep;23(7):2089-101. doi: 10.1007/s11136-014-0658-z. Epub 2014 Mar 7.

- Anota A, Bascoul-Mollevi C, Conroy T, Guillemin F, Velten M, Jolly D, Mercier M, Causeret S, Cuisenier J, Graesslin O, Hamidou Z, Bonnetain F. Item Response Theory and Factor Analysis as a mean to characterize occurrence of Response Shift in a longitudinal quality of life study in breast cancer patients. Health and Quality of Life Outcomes. 2014 Mar 8;12:32. doi: 10.1186/1477-7525-12-32.

- Bonnetain F, Fiteni F, Anota A. Statistical Challenges in the Analysis of the Health-related Quality of Life in Cancer Clinical Trials. **Under Review in *Journal of Clinical Oncology***

- Anota A, Savina M, Bascoul-Mollevi C, Bonnetain F. QoLR: an R Package for the Longitudinal Analysis of Health-Related Quality of Life in Oncology. **Under Review in *Journal of Statistical Software***

- Anota A, Mouillet G, Trouilloud I, Dupont-Gossart A.C., Artru P, Lecomte T, Zaanan A, Gauthier M, Fein F, Dubreuil O, Paget-Bailly S, Taieb J, Bonnetain F. Sequential FOLFIRI.3 + Gemcitabine improves health-related quality of life deterioration-free survival of patients with metastatic pancreatic adenocarcinoma: a randomized phase II trial. **Under Review in *PLOS ONE***

- Anota A, Barbieri A, Savina M, Pam A, Gourgou-Bourgade S, Bonnetain F, Bascoul-Mollevi C. Comparison of three longitudinal analysis models for the health-related quality of life in oncology: a simulation study. **Soumis à *Health and Quality of Life Outcomes***

- Anota A, Boulin M, Dabakuyo S, Hillon P, Cercueil JP, Minello A, Jouve JL, Paoletti X, Bedenne L, Guiu B, Bonnetain F. An explorative study to assess added value of the health-related quality of life in a phase I trial: Idarubicin-loaded beads for chemoembolization of hepatocellular carcinoma. **Article en cours de rédaction**

- Anota A, Bascoul-Mollevi C, Conroy T, Guillemin F, Velten M, Jolly D, Hamidou Z, Bonnetain F. Is Structural Equation Modeling a valuable tool to highlight the occurrence of Response Shift Effect with EORTC health-related quality of life questionnaires? **Article en cours de rédaction**

Communications orales

- **Internationales**

- A. Anota, F. Bonnetain. Time to Health-related Quality of Life score deterioration as a modality of longitudinal analysis in Quality of Life studies in oncology integrating the occurrence of a response shift effect. International workshop « Response Shift and subjective measures in health science », 7 juin 2013, Nantes

- A. Anota, C. Bascoul-Mollevi, T. Conroy, F. Guillemin, M. Velten, D. Jolly, A. Pam, F. Bonnetain. Structural Equation Modeling to characterize the occurrence of the Response Shift effect in a longitudinal quality of life study. 21th Annual Conference of the International Society for Quality of Life Research (ISOQOL), 15-18 octobre 2014, Berlin (Allemagne)

- **Nationales**

- A. Anota, C. Bascoul-Mollevi, F. Guillemin, T. Conroy, M. Velten, D. Joly, M. Mercier, S. Causeret, S. Dabakuyo, F. Bonnetain. Modèles d'« Item Response Theory » et analyses factorielles pour caractériser l'occurrence de la « Response shift » dans l'étude longitudinale de la qualité de vie de patientes atteintes de cancer du sein. EPICLIN 6/ 19èmes Journées des Statisticiens de CLCC, 9-11 mai 2012, Lyon

- A. Anota, F. Bonnetain. Analyse longitudinale de la Qualité de Vie relative à la santé : « Response Shift » et méthodologie d'analyse. Journée scientifique annuelle de la SFR FED4234, 24 mai 2013, Besançon

- A. Anota, M. Savina, C. Bascoul-Mollevi, F. Bonnetain. QoLR : un package R pour l'analyse longitudinale de la qualité de vie en cancérologie. 20èmes Journées des Statisticiens des CLCC, 6-7 juin 2013, Marseille

- A. Anota, I. Trouilloud, M. Gauthier, J. Taieb, S. Paget Bailly, D. Vernerey, F. Bonnetain. Analyse longitudinale de la qualité de vie de patients atteints d'un cancer du pancréas métastatique non pré-traité par la méthode du temps jusqu'à détérioration tenant compte du profil des données manquantes par le score de propension. 20èmes Journées des Statisticiens des CLCC, 6-7 juin 2013, Marseille

- A. Anota, C. Bascoul-Mollevi, F. Bonnetain. Comparaison de modèles longitudinaux pour l'analyse de la qualité de vie relative à la santé en cancérologie. Workshop national « Quels apports des mathématiques à la mesure de la qualité de vie en oncologie », 12-13 septembre 2013, Marseille

- A. Anota, A. Barbieri, M. Savina, A. Pam, S. Gourgou-Bourgade, F. Bonnetain, C. Bascoul-Mollevi. Comparison of three longitudinal analysis models for the health-related quality of life in oncology: a simulation study. Workshop « Evaluation et analyse de la qualité de vie : nouveaux développements méthodologiques », 3-4 avril 2014, Montpellier

- A. Anota, C. Bascoul-Mollevi, T. Conroy, F. Guillemin, M. Velten, D. Jolly, A. Pam, F. Bonnetain. Modèles à équations structurelles pour caractériser l'occurrence de la Response Shift dans l'évaluation longitudinale de la qualité de vie relative à la santé. Workshop « Evaluation et analyse de la qualité de vie : nouveaux développements méthodologiques », 3-4 avril 2014, Montpellier

- A. Anota, A. Barbieri, M. Savina, A. Pam, S. Gourgou-Bourgade, F. Bonnetain, C. Bascoul-Mollevi. Comparison of three longitudinal analysis models for the health-related quality of life in oncology: a simulation study. EPICLIN 8/21èmes Journées des Statisticiens des CLCC, 14-16 mai 2014, Bordeaux

Communications affichées

- **Internationales**

- A. Anota, C. Bascoul-Mollevi, F. Guillemin, T. Conroy, M. Velten, D. Joly, M. Mercier, O. Graesslin, S. Causeret, J. Cuisenier, S. Dabakuyo, F. Bonnetain. Item Response Theory and Factor Analysis as mean to characterize occurrence of Response Shift for longitudinal quality of life study in breast cancer patients. ESMO 2012 congress, 29 septembre - 2 octobre 2012, Vienne (Autriche)

- A. Anota, C. Bascoul-Mollevi, F. Guillemin, T. Conroy, M. Velten, D. Joly, M. Mercier, O. Graesslin, S. Causeret, J. Cuisenier, S. Dabakuyo, F. Bonnetain. Item Response Theory and Factor Analysis as mean to characterize occurrence of Response Shift for longitudinal quality of life study in breast cancer patients. 19th Annual Conference of the International Society for Quality of Life Research (ISOQOL), 24-27 octobre 2012, Budapest (Hongrie)

- A. Anota, I. Trouilloud, M. Gauthier, J. Taieb, F. Bonnetain. Time to health-related Quality of Life score deterioration as a modality of longitudinal analysis using inverse probability weighting to deal with missing data: a phase II trial in patients with metastatic non pre-treated pancreatic adenocarcinoma. European Cancer Congress, 27 septembre -1er octobre 2013, Amsterdam (Pays-Bas)

- A. Anota, M. Boulin, S. Dabakuyo, L. Bedenne, B. Guiu, F. Bonnetain. Exploratory study of health-related quality of life in a phase I trial studying Idarubicin-loaded beads for chimioembolization of hepatocellular carcinoma. European Cancer Congress, 27 septembre - 1er octobre 2013, Amsterdam (Pays-Bas)

- A. Anota, M. Savina, C. Bascoul-Mollevi, F. Bonnetain. QoLR: an R package for the longitudinal analysis of Health-related quality of life in oncology. 20th Annual Conference of the International Society for Quality of Life Research (ISOQOL), 9-12 octobre 2013, Miami

(Etats-Unis) *(Prix du meilleur poster)*

- A. Anota, I. Trouilloud, M. Gauthier, J. Taieb, F. Bonnetain. Time to health-related Quality of Life score deterioration as a modality of longitudinal analysis using inverse probability weighting to deal with missing data: a phase II trial in patients with metastatic non pre-treated pancreatic adenocarcinoma. 20th Annual Conference of the International Society for Quality of Life Research (ISOQOL), 9-12 octobre 2013, Miami (Etats-Unis)

- A. Anota, A. Barbieri, M. Savina, A. Pam, M. Ychou, S. Gourgou-Bourgade, F. Bonnetain, C. Bascoul-Mollevi. Comparison of three longitudinal analysis models for the health-related quality of life in oncology: a simulation study. ISOQOL, 21th Annual Conference of the International Society for Quality of Life Research (ISOQOL), 15-18 octobre 2014, Berlin (Allemagne)

➢ **Nationales**

- A. Anota, M. Savina, C. Bascoul-Mollevi, F. Bonnetain. QoLR: un package R pour l'analyse longitudinale de la qualité de vie en cancérologie, EPICLIN 7 (Epidémiologie et recherche clinique), 16-17 mai 2013, Paris

- A. Anota, C, Bascoul-Mollevi, F. Bonnetain. Evaluation et comparaison de trois méthodes statistiques pour l'analyse longitudinale de la qualité de vie relative à la santé en cancérologie. 7ème Forum du Cancéropôle Grand-Est, 25-26 novembre 2013, Strasbourg

- A. Anota, M. Boulin, S. Dabakuyo, L. Bedenne, B. Guiu, F. Bonnetain. Analyse exploratoire de la qualité de vie relative à la santé dans un essai de phase I étudiant l'impact de l'administration d'une dose d'idarubicine associée à une chimioembolisation des carcinomes hépatocellulaires non résécables. 7ème Forum du Cancéropôle Grand-Est, 25-26 novembre 2013, Strasbourg

- A. Anota, C. Bascoul-Mollevi, T. Conroy, F. Guillemin, M. Velten, D. Jolly, A. Pam, F. Bonnetain. Modèles à équations structurelles pour caractériser l'occurrence de la Response Shift dans l'évaluation longitudinale de la qualité de vie relative à la santé. EPICLIN 8/21èmes Journées des Statisticiens des CLCC, 14-16 mai 2014, Bordeaux *(Prix du meilleur poster)*

- A. Anota, C. Bascoul-Mollevi, F. Bonnetain. Analyse longitudinale de la qualité de vie relative à la santé. Congrès ADELF, 11-13 septembre 2014, Nice

I. INTRODUCTION

A ce jour, la survie globale est considérée par la Food and Drug Administration (FDA) comme le gold standard en tant que critère de jugement principal dans les essais de phase III en cancérologie pour pouvoir évaluer le bénéfice clinique de nouvelles stratégies thérapeutiques (Food & Administration, 2007; Peppercorn *et al*, 2011). Cependant, en raison de la multiplicité des nouveaux traitements efficaces pour la majorité des situations thérapeutiques, il est souvent nécessaire d'augmenter la taille de l'échantillon ainsi que la durée de l'étude afin d'observer un nombre de décès suffisant pour pouvoir conclure à la supériorité ou non du nouveau traitement comparativement au traitement de référence.
Ainsi, des critères de jugement alternatifs sont de plus en plus utilisés comme des critères centrés sur la tumeur, en particulier la survie sans progression (Fiteni *et al*, 2014). Ces critères ont l'avantage de pouvoir être évalués plus précocement que la survie globale (Fiteni *et al*, 2014). Ils permettent de diminuer le nombre de sujets nécessaires ainsi que la durée de l'essai. Or, pour pouvoir être utilisés en tant que critère de jugement principal, il faut soit avoir validé ces critères en tant que critère de jugement substitutif de la survie globale, soit s'assurer du bénéfice clinique pour le patient avec l'utilisation conjointe d'un critère centré sur le patient. Ainsi, la qualité de vie relative à la santé (QdV) a souvent été proposée comme critère de jugement secondaire afin de s'assurer du bénéfice clinique.

Bien que la FDA et l' « American Society of Clinical Oncology » (ASCO) considèrent la QdV comme second critère de jugement principal en l'absence d'effet sur la survie globale (Beitz *et al*, 1996), les résultats des études de QdV restent encore peu utilisés en pratique clinique. Par ailleurs, encore peu d'essais cliniques considèrent à ce jour la QdV comme co-critère de jugement principal avec la survie sans progression. Cette sous-exploitation de la QdV est en particulier due à la complexité des données de QdV.

En effet, la QdV est d'une part un critère de jugement subjectif puisque chaque patient à sa propre définition de la QdV. Elle est considérée comme un trait latent, i.e. qu'on ne peut ni l'observer ni la quantifier directement. Sa mesure se fait donc principalement de façon indirecte par le biais d'auto-questionnaires administrés aux patients. Cette auto-évaluation est sujette aux données manquantes puisque le patient ne répond pas nécessairement à l'ensemble des questions ainsi qu'à chaque questionnaire planifié dans le design de l'étude. Ces données manquantes peuvent biaiser l'analyse puisqu'elles peuvent dépendre de l'état de santé et/ou

de la QdV du patient. D'autre part, le caractère subjectif de la QdV induit également la notion de dynamique. En effet, la définition de la QdV est susceptible d'être modifiée au cours du temps pour un même patient en raison de l'apparition d'un changement de l'état de santé de l'individu pouvant affecter sa vie quotidienne. Celui-ci peut alors ajuster ses références internes et revoir à la baisse son niveau d'exigences vis-à-vis de sa QdV. L'individu n'évalue donc pas nécessairement sa QdV avec les mêmes critères au cours du temps. Cette dynamique, reflétée par un effet « Response Shift », peut biaiser l'analyse longitudinale de la QdV si elle n'est pas prise en compte de façon adéquate soit dans le design de l'étude, soit dans la méthode d'analyse.

L'analyse longitudinale de la QdV doit donc être réalisée selon une méthodologie rigoureuse permettant de prendre en compte ces différents facteurs. Il parait également important de pouvoir proposer des méthodes statistiques permettant d'obtenir des résultats accessibles et compréhensibles par les cliniciens. En raison de ces différents facteurs pouvant complexifier l'analyse de la QdV, il n'existe pas à ce jour de recommandation pour l'analyse longitudinale de la QdV.

L'objectif de notre travail est de faire le point sur ces facteurs limitants et de proposer des méthodes adéquates pour une interprétation robuste des données de QdV longitudinales.

Un premier chapitre de ce travail rappelle le contexte concernant la QdV, tel que sa définition, sa mesure et les méthodes utilisées en pratique pour son analyse longitudinale dans les essais cliniques en cancérologie.

Ensuite, les objectifs de mon travail de thèse sont présentés ainsi que les différents travaux réalisés au cours de mon doctorat. Ces travaux sont centrés sur la méthode du temps jusqu'à détérioration d'un score de QdV, en tant que modalité d'analyse longitudinale dans les essais cliniques en cancérologie, ainsi que sur la caractérisation de l'occurrence de l'effet Response Shift.

Enfin, les deux derniers chapitres sont consacrés à la discussion générale des différents résultats et de nos perspectives de recherche et à la conclusion.

II. CONTEXTE

1. La qualité de vie relative à la santé

1.1. Définition et caractère multidimensionnel

En 1948, l'Organisation Mondiale de la Santé (OMS) a défini la santé comme « un état de complet bien-être physique, mental et social, et pas seulement l'absence de maladie ou d'infirmité » (WHO, 1948).

En 1993, l'OMS définie la qualité de vie comme « la perception qu'a un individu de sa place dans l'existence, dans le contexte de la culture et du système de valeurs dans lequel il vit, en relation avec ses objectifs, ses attentes, ses normes et ses inquiétudes. Il s'agit donc d'un large champ conceptuel, englobant de manière complexe la santé physique de la personne, son état psychologique, son niveau d'indépendance, ses relations sociales, ses croyances personnelles et sa relation avec les spécificités de son environnement » (WHOQOL, 1993).

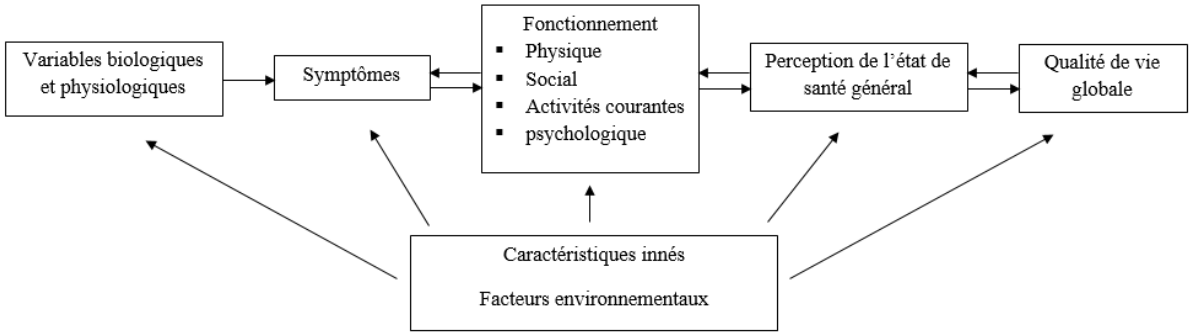
La qualité de vie relative à la santé (QdV) découle de cette définition et intègre l'impact de la maladie et du traitement sur la qualité de vie du patient. Certaines conséquences indirectes de la maladie telles que la perte d'un emploi ou les difficultés financières sont également prises en compte.

Bien qu'il n'existe pas de consensus autour de la définition de la QdV, elle est généralement considérée comme un concept multidimensionnel qui inclut au minimum le bien-être physique, psychologique et social mais aussi les symptômes liés à la maladie et aux traitements.

La QdV entre dans le champ des « Patient-Reported Outcomes » (PROs), i.e. des mesures de l'état de santé perçue par le patient (Doward & McKenna, 2004; Fayers & Machin, 2007). Ces mesures doivent donc être rapportées par le patient lui-même. Les PROs peuvent correspondre à une large variété de paramètres comme les symptômes liés à la maladie ou au traitement, par exemple, la fatigue ou la douleur. La satisfaction vis-à-vis des soins entre également dans le champ des PROs.

L'évaluation de la QdV est influencée par un certain nombre de paramètres. Bien qu'il n'existe pas de consensus autour de la définition de la QdV, un modèle conceptuel proposé par Wilson et Cleary en 1995 a souvent servi d'illustration pour décrire les différents facteurs

influençant potentiellement l'évaluation de la QdV et les différentes interactions possibles avec des paramètres extérieurs (Wilson & Cleary, 1995). Le modèle ainsi décrit est un modèle en cinq niveaux. Le premier niveau correspond aux critères biologiques et physiologiques. Ceux-ci ont un impact sur le niveau symptomatique du patient, caractérisant la perception que le patient a de ses symptômes. Le troisième niveau correspond à la perception que le patient a de son état fonctionnel, incluant son état physique et émotionnel. Ce troisième niveau est en lien direct avec la perception que le patient a de son état de santé globale correspondant au quatrième niveau. Cette perception de sa santé globale va enfin jouer un rôle direct sur son évaluation de sa QdV globale. Dans ce modèle, le lien entre chaque niveau est unidirectionnel. En 2007, Osoba et al. (Osoba, 2007) ont proposé une version révisée de ce schéma dans laquelle des interactions dans les deux sens sont possibles entre l'état symptomatique, l'état fonctionnel, la santé perçue et la qualité de vie globale (Figure 1).



<u>Figure 1</u> : Modification du modèle de Wilson et Cleary indiquant le potentiel d'une interaction à double sens entre plusieurs composantes du modèle (Osoba, 2007)

1.2. <u>Mesure subjective</u>

La QdV est un trait latent, i.e. qu'on ne peut l'observer directement. On cherche donc à approcher sa mesure soit directement et d'un point de vue qualitatif, par des entretiens ouverts ou semi-directifs avec un psychologue, soit indirectement et d'un point de vue quantitatif, par le moyen de questionnaires administrés au patient (Fayers & Machin, 2007).

La QdV est une mesure subjective puisque chaque individu a sa propre définition de la QdV. Lorsqu'un patient évalue son niveau de QdV, son jugement dépend de ses références internes, de l'importance relative qu'il accorde aux différentes dimensions de la QdV et de sa propre définition de la QdV. L'évaluation de la QdV est donc influencée par les attentes et

espérances de santé du patient (Bullinger, 2002; Wiklund, 2004). A titre d'exemple, un individu avec une maladie chronique n'aura pas les mêmes attentes et espérances de santé qu'un individu habituellement en bonne santé et qui vient d'être diagnostiqué pour un cancer.

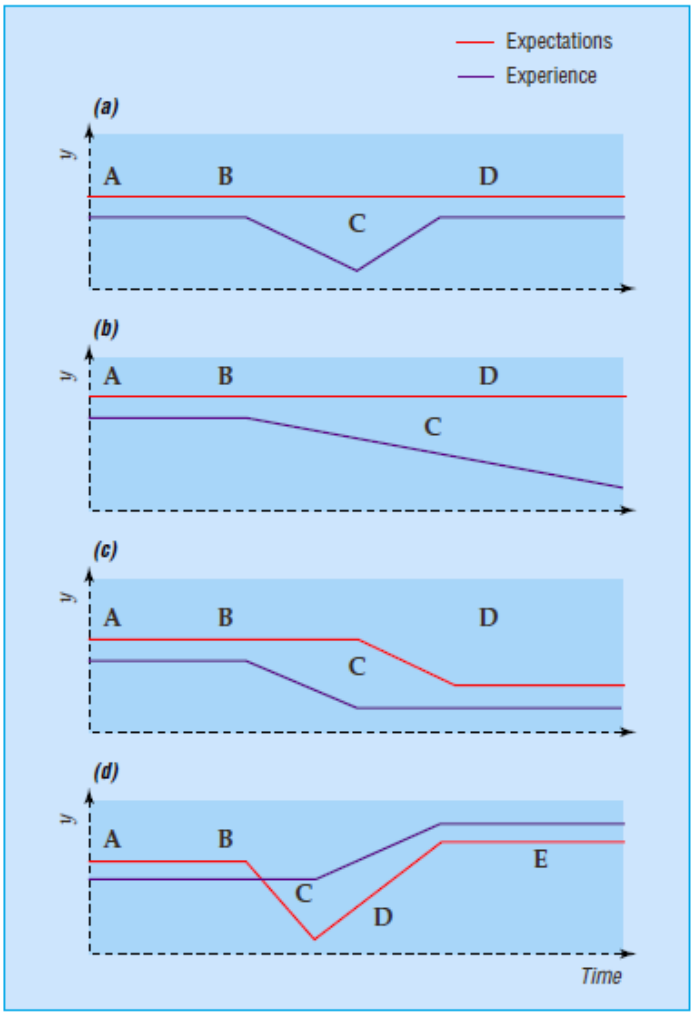
La nature subjective de la mesure de la QdV est considérée dogmatiquement comme un frein à son utilisation et à la prise en compte de ses résultats par des professionnels de santé. Ces derniers sont davantage attachés à des critères dont l'évaluation est plus objective comme la survie globale. Pourtant, certains critères tumoraux comme la réponse tumorale peuvent également être mesurés selon une certaine imprécision ou subjectivité.

Une mesure de la QdV par une tierce personne comme un médecin pourrait sembler plus objective. Cependant, cette subjectivité est une part inhérente à la mesure de la QdV puisque seul le patient est capable d'évaluer sa propre perception de son état de santé ou de sa QdV. En outre, il n'existe pas de concordance entre la mesure effectuée par le patient lui-même et celle effectuée par un médecin ou un proche, en particulier pour les dimensions émotionnelles et psychologiques (Moinpour *et al*, 2000; Stephens *et al*, 1997). Seuls les critères les plus objectifs de la QdV tels que les symptômes physiques et l'état fonctionnel présentent une meilleure concordance (Sprangers & Aaronson, 1992) De plus, une évaluation effectuée par un médecin ou un proche peut également être influencée par leurs propres perceptions et espérances de santé (Addington-Hall & Kalra, 2001).

1.3. Concept dynamique

La QdV est également un concept dynamique. En effet, un individu donné n'évalue pas nécessairement sa QdV selon les mêmes critères au cours du temps, ses attentes et espérances de santé peuvent être modifiées du fait du diagnostic d'une maladie tel qu'un cancer. Le patient peut ainsi s'adapter à la maladie et à la toxicité du traitement et revoir, par exemple, ses espérances de santé à la baisse. Le patient peut également accorder moins d'importance à son état physique et être plus attaché et plus proche de sa famille et de ses amis qu'il ne l'était avant l'apparition de la maladie. Cet aspect dynamique est illustré par le modèle interactif de Wilson et Cleary où la perception de la santé et de la QdV du patient est entre autre influencée par des facteurs extérieurs tel que l'échelle de valeurs des domaines de QdV (Wilson & Cleary, 1995). Ce processus dynamique est également reflété par un effet « Response Shift » (Gibbons, 1999; Howard *et al*, 1979a; Howard *et al*, 1979b).

Carr *et al.* illustrent la complexité et la dynamique de la relation entre les attentes et expériences de santé et la perception du patient de son état de santé et de sa QdV (Carr *et al*, 2001). Selon les auteurs, les individus évaluent leur QdV en comparant leurs attentes avec leurs expériences. La Figure 2 représente quatre situations différentes de douleurs au niveau du dos survenant chez une femme d'âge moyen et son évolution au cours du temps. Pour chacune de ces situations, on représente la relation entre les attentes de santé (« expectations ») et l'expérience de la patiente.



<u>Figure 2</u> : Quatre modèles de douleurs survenant chez un individu à différents moments (A-E) (source Carr et al. (Carr *et al*, 2001)) selon divers scénarios : (a) épisode aigu, (b) épisode chronique, (c) acceptation du patient de l'épisode chronique, (d) variation des effets des expériences et attentes au cours du temps.

Ainsi, la situation (a) représente un épisode aigu, la douleur est donc passagère et les espérances de santé de la patiente reste invariante au cours du temps. La situation (b) représente un épisode chronique où la douleur ressentie par la patiente au niveau du dos s'accentue au cours du temps. Dans cette situation, la femme garde l'espoir de voir son état de

santé s'améliorer. Comme ses attentes de santé restent inchangées, cela aura un impact dans sa capacité à travailler ou dans ses relations sociales. Dans la situation (c), la patiente accepte cette douleur et revoit ainsi ses attentes de santé à la baisse. Enfin, dans la situation (d), il s'agit d'une situation où la patiente souffre de douleurs pendant une longue période (période A). Cette patiente s'adapte à son état de santé et ses attentes de santé deviennent alors plus faibles. Elle peut suivre alors un programme de gestion de la douleur qui lui permet de contrôler sa douleur (période C). Pendant cette période, son expérience est plus élevée que ses attentes. Elle revoit alors ses attentes à la hausse (période D) et on obtient finalement une stabilisation à un niveau plus élevé qu'initialement (période E).

Ces modifications d'attentes et espérances de santé du patient sont les conséquences de l'existence de stratégies de coping du patient, i.e. de stratégies d'adaptation du patient vis-à-vis de la maladie (Dunkel-Schetter *et al*, 1992). Le coping correspond à l'ensemble des efforts cognitifs et comportementaux fournies par le patient dans le but de maîtriser l'impact de la maladie sur sa QdV. Différents profils de coping existent et peuvent être identifiés tels que « se focaliser sur les évènements positifs », « rechercher un soutien social », « prendre de la distance », etc. (Dunkel-Schetter *et al*, 1992).

2. Critères de jugements dans les essais cliniques en cancérologie

Les critères de jugements font référence à des mesures cliniques et biologiques pour évaluer l'efficacité d'une stratégie thérapeutique. Dans les essais cliniques en cancérologie, les critères de jugements peuvent être regroupés en deux catégories : les critères de jugements « centrés sur le patient » comme la survie globale et la QdV et les critères de jugements « centrés sur la tumeur » comme la survie sans progression (Bonnetain, 2010; Fiteni *et al*, 2014).

A ce jour, la survie globale reste le « gold standard » pour évaluer l'efficacité d'une prise en charge. Néanmoins, compte tenu du nombre croissant de traitements efficaces pour une grande majorité des cancers, il est nécessaire d'augmenter le nombre de patients à inclure ainsi que la durée de suivi de chaque patient afin d'observer un nombre de décès suffisant, i.e. une puissance statistique suffisante (Fiteni *et al*, 2014).

Ainsi, de plus en plus de chercheurs se tournent vers des paramètres centrés sur la tumeur telle que la survie sans progression. Ces paramètres peuvent être évalués plus précocement que la survie globale. De plus, si ces critères sont validés comme critères substitutifs à la survie globale, ils peuvent être considérés comme critères cliniques. Cependant, la survie sans progression a rarement été validée comme critère substitutif à la survie globale. A titre d'exemple, cette validation a déjà été réalisée en situation avancée du cancer du sein (Beauchemin *et al*, 2014), du mélanome (Kim, 2014) et du carcinome rénal (Negrier *et al*, 2014). Pour pouvoir considérer la survie sans progression comme critère de jugement principal, elle doit avoir été validée en tant que critère substitutif de la survie globale pour la localisation cancéreuse et le type de traitement considéré. La multiplicité des définitions possibles pour les évènements à prendre en considération est également une limite de ces critères (Bellera *et al*, 2013). Dans ce contexte, il parait indispensable de s'assurer du bénéfice clinique pour le patient. La QdV a ainsi été reconnue comme second critère de jugement principal par l'ASCO et la FDA en l'absence d'effet sur la survie globale (Beitz *et al*, 1996).

La discussion actuelle serait donc de considérer la QdV comme co-critère de jugement principal ou critère de jugement composite avec un critère tumoral comme la survie sans progression (Amir *et al*, 2012; Bonnetain *et al*, 2012). A titre d'illustration, un essai clinique de phase III a récemment investigué l'impact de l'ajout du bévacizumab à un traitement standard chez des patients atteints d'un glioblastome nouvellement diagnostiqué (Chinot *et al*, 2014). Dans cet essai, la survie sans progression et la survie globale étaient considérées comme co-critères de jugement principal. La QdV était un critère de jugement secondaire. Cet essai a montré un bénéfice du bévacizumab au niveau de la survie sans progression et une absence d'effet sur la survie globale. La survie sans progression n'ayant pas été validé en tant que critère de jugement substitutif de la survie globale, l'étude de la QdV est indispensable pour s'assurer du bénéfice clinique pour le patient. L'étude de la QdV a montré que les patients du groupe bevacizumab présentaient une détérioration plus tardive de la QdV contrairement aux patients du groupe standard. Les résultats de QdV étaient donc, dans cet essai, en cohérence avec les résultats de la survie sans progression.

3. Les instruments de mesure de la qualité de vie

Généralement, la QdV est mesurée par le biais de questionnaires administrés aux patients (Fayers & Machin, 2007). De nombreux questionnaires de QdV et de PROs ont ainsi été développés. Ces questionnaires doivent respecter un certain nombre de propriétés psychométriques pour pouvoir être validés et utilisés avec fiabilité. Ils doivent être adaptés à la population d'étude et doivent être validés dans la langue d'utilisation.

Ces questionnaires peuvent être génériques ou bien spécifiques d'une maladie. Les questionnaires génériques peuvent être administrés à tout individu quel que soit son état de santé. Les questionnaires spécifiques d'une maladie ont en revanche été développés pour une pathologie donnée et mesurent ainsi des domaines de QdV potentiellement impactés par cette pathologie.

3.1. Items et scores

La QdV étant un concept multidimensionnel, la majorité des questionnaires de QdV sont des questionnaires multidimensionnels. Chaque dimension (ou échelle de QdV) est évaluée par une ou plusieurs questions appelées items. Les modalités de réponse aux items peuvent être de type dichotomiques (réponse de type « Oui » vs. « Non ») ou polytomiques. Les items polytomiques peuvent correspondre à une échelle ordinale ou nominale. Les instruments de mesure de QdV spécifiques du cancer contiennent généralement des items polytomiques ordinaux. Les échelles de Likert pour lesquelles chaque modalité de réponse i est accompagnée par une étiquette décrivant l'état de santé associé sont particulièrement utilisées. Par exemple, un certain nombre de questionnaires utilisent une échelle de Likert à quatre modalités de réponse du type : 1 « Pas du tout » / 2 « Un peu » / 3 « Assez » / 4 « Beaucoup » (Aaronson *et al*, 1993).

Pour chaque dimension, un score est généré à partir des réponses aux items et reflétant le niveau de l'individu pour la dimension concernée. Il s'agit généralement de la moyenne des réponses aux items. Si la dimension est évaluée à la fois par des items positifs et négatifs (i.e. où une réponse élevée correspondra à un niveau de QdV élevé ou faible selon l'item), il est

nécessaire d'inverser certains items tel qu'un score élevé correspondra à un niveau de QdV élevé ou faible.

Selon les questionnaires, un pourcentage de données manquantes peut être toléré afin de pouvoir estimer le score d'un patient même si celui-ci n'a pas répondu à l'ensemble des items. Le score peut alors être calculé à partir des items répondus, en considérant, par exemple, que les items non renseignés ne diffèrent pas significativement des items répondus. Ce score est généralement standardisé de 0 à 100 afin de faciliter la comparaison ente les dimensions et entre les questionnaires. Pour les questionnaires multidimensionnels, des scores résumés peuvent être proposés, afin de refléter un niveau global de QdV par exemple, et correspondant à la moyenne obtenue pour l'ensemble ou un sous-ensemble de dimensions. Les algorithmes de calcul des scores ainsi que le pourcentage de données manquantes toléré sont propres à chaque questionnaire.

3.2. Echelle uni ou multi-items

Chaque dimension de la QdV peut être évaluée par un item (échelle uni-item) ou plusieurs items (échelle multi-items). La QdV étant un concept complexe, il peut être difficile de l'évaluer par une seule question globale, et dans ce cas une échelle multi-items semble préférable. Les échelles multi-items permettent également d'évaluer plus précisément le domaine de QdV évalué que les échelles uni-items.

Certains questionnaires peuvent mélanger des échelles uni et multi-items (Aaronson *et al*, 1993), ne contenir que des échelles multi-items (Cella *et al*, 1993) ou uniquement des échelles uni-items (Selby *et al*, 1984). Ainsi, pour certaines dimensions assez concrètes et objectives telles que des symptômes de diarrhée, de constipation ou de perte d'appétit, une échelle uni-item parait pertinente. En revanche, pour des dimensions plus difficile à définir, tel que l'état émotionnel, une échelle multi-item serait plus adaptée afin de capter au mieux l'ensemble des facteurs à prendre en considération pour évaluer cette dimension.

3.3. Echelle visuelle analogique

La mesure de la QdV globale ou d'un symptôme comme la fatigue ou la douleur peut également être réalisée à l'aide d'une échelle visuelle analogique (EuroQol, 1990; Selby *et al*, 1984). Il est ainsi demandé au patient d'évaluer sur une échelle graduée ou une ligne de 0 à 10 ou de 0 à 100 son état de santé global, sa QdV globale, son niveau de fatigue ou de douleur par exemple. Chaque extrémité de l'échelle est labellisée d'une étiquette qualifiant l'état correspondant (par exemple: « Pire état de santé possible »; «Meilleur état de santé possible »). Les patients doivent alors marquer la ligne à une certaine distance entre les deux extrémités correspondant au mieux à leur état de santé. Un attaché de recherche clinique doit alors relever la mesure rapportée par le patient en évaluant la distance entre l'extrémité inférieure de l'échelle et la valeur rapportée par le patient à l'aide d'une règle graduée. Pour une telle mesure, aucun score n'a besoin d'être calculé puisque l'analyse peut se faire directement avec la valeur relevée. Cette mesure pourrait avoir l'avantage d'être plus précise et plus proche d'une mesure quantitative que le sera un score généré d'après les réponses aux items qui le constituent. Néanmoins, certaines études ont montré que les patients avaient plutôt tendance à choisir les niveaux extrêmes ou moyens (McCormack *et al*, 1988). Cette méthode requiert en outre (tout comme les autres échelles uni-items) que le patient définisse lui-même le domaine évalué et prenne en considération tous les facteurs pouvant influencer ce domaine et leur importance relative (Fayers & Machin, 2007).

3.4. Auto-évaluation, hétéro-évaluation ou proxy

La QdV étant un critère subjectif, une évaluation de la QdV par le patient lui-même est préférable (Addington-Hall & Kalra, 2001). D'autre part, il est souhaitable de procéder à une auto-évaluation, i.e. que le patient remplisse lui-même le questionnaire, sans qu'il ne lui soit énoncé par une tierce personne. Ainsi, le patient ne sera pas influencé dans ses réponses par la personne lui énonçant le questionnaire. Si le questionnaire lui est énoncé par son médecin, le patient peut être sujet à un biais de désirabilité sociale. Si le questionnaire lui est énoncé par un proche, le patient peut ne pas se sentir libre dans son choix de réponse, avoir peur du jugement, en particulier pour des questions relatives à la famille, à certains symptômes dérangeants ou à la sexualité.

Cependant, si le patient est trop fatigué pour remplir le questionnaire lui-même, il peut dans certains cas lui être énoncé par un attaché de recherche clinique, un médecin ou un proche. On parle alors d'une hétéro-évaluation. Il existe également des hétéro-questionnaires développés pour être énoncés au patient par une tierce personne. Ces questionnaires sont particulièrement utilisés pour les patients en fin de vie trop faibles ou fatigués pour remplir eux-mêmes le questionnaire.

Enfin, dans certaines situations, une évaluation par le patient peut ne pas être réalisable. Ainsi, des instruments de mesures de la QdV ont été développés pour être renseignés par le médecin ou un proche du patient : il s'agit de mesure proxy (Pickard & Knight, 2005). Cette tierce personne évalue alors elle-même la QdV du patient. Ce type de questionnaire est employé uniquement dans certaines situations où le patient est incapable de donner une réponse cohérente: par exemple, si le patient est trop jeune, trop âgé, trop malade ou présente une déficience mentale.

3.5. Les questionnaires génériques

Les questionnaires génériques sont conçus pour pouvoir être administrés à tout type d'individu quel que soit leur état de santé. Ils permettent ainsi de comparer des populations très hétérogènes et des patients avec des pathologies très diverses. De nombreux questionnaires génériques ont été développés.

3.5.1. Questionnaire MOS SF-36

Le questionnaire le plus utilisé est le questionnaire « Medical Outcome Study Short Form-36 items » (MOS SF-36) (Ware & Sherbourne, 1992). Ce questionnaire contient 36 items et permet d'évaluer 8 dimensions de QdV : le fonctionnement physique, le rôle physique, les douleurs physiques, l'état de santé général, la vitalité, le fonctionnement social, le rôle-émotionnel et la santé mentale. Les items de ce questionnaire sont construits sur une échelle de Likert à 3 ou 5 modalités de réponses selon les items. Chaque score est standardisé de 0 à 100 tel que 100 représente un bon niveau de QdV, i.e. un haut niveau fonctionnel ou un bas niveau symptomatique selon l'échelle. De plus, deux scores résumés sont proposés et correspondant à la santé physique générale et à la santé mentale générale. Le score global de

santé physique est évalué par les dimensions fonctionnement physique, rôle physique, douleurs physiques, et état de santé général. Le score global de santé mentale est évalué par la vitalité, le fonctionnement social, le rôle-émotionnel et la santé mentale. Ce questionnaire contient également une question de transition concernant l'état de santé général du patient : « Comparé à il y a un an, comment évaluerez-vous votre état de santé actuel ? ».

3.5.2. Questionnaires WHOQOL

Le questionnaire WHOQOL (« World Health Organization Quality Of Life ») est un questionnaire générique développé par l'OMS (WHOQOL, 1993). Deux versions de ce questionnaire existent : le WHOQOL-100 contenant 100 items et le WHOQOL-BREF correspondant à une version plus courte à 26 items. Ces questionnaires étant issus d'une collaboration internationale, ils seraient peu sensibles à un biais culturel. Ils permettent d'explorer la santé physique et psychique, les relations sociales, l'environnement, le niveau d'indépendance du patient mais aussi la spiritualité et les croyances du patient.

3.5.3. Questionnaire EUROQoL EQ-5D

Le questionnaire EUROQOL EQ-5D est également un questionnaire générique (EuroQol, 1990). Ce questionnaire contient 5 items mesurant 5 dimensions de QdV : la mobilité, l'autonomie de la personne, les activités usuelles, les douleurs et inconforts ainsi que l'état d'anxiété et de dépression. Ces items sont construit sur une échelle de Likert à 3 modalités de réponses codées : 1 « pas de problème » / 2 « problèmes modérés » / 3 « problèmes sévères ». Le manuel d'utilisation du questionnaire EQ-5D propose de calculer un code résumé pour chaque patient (appelé « digit code ») correspondant aux réponses données aux 5 items et reflétant l'état de santé du patient (Cheung *et al*, 2009). Par exemple, l'état « 11111 » indique aucun problème dans les 5 domaines de QdV évalués ; l'état « 11223 » indique aucun problème de mobilité ni de l'autonomie, quelques difficultés au niveau des activités usuelles, des douleurs et inconforts modérés et un état d'anxiété et de dépression élevé. Ainsi, un total de 243 états possibles sont définis de la même façon. Un score index est alors proposé. Ce score est souvent utilisé dans les analyses de QALYs (voir paragraphe 6.2.6). Une analyse descriptive de l'état de santé de la population est généralement réalisée d'après les items polytomiques originaux ou dichotomisés. Ce questionnaire contient également une échelle

visuelle analogique correspondant à une règle graduée de 0 à 100. Il est demandé au patient d'évaluer sur cette échelle son état de santé général actuel où 0 correspond au pire et 100 au meilleur état de santé possible.

3.5.4. Questionnaire SEIQoL

Le questionnaire SEIQoL (« Schedule for the Evaluation of Individual Quality of Life ») est un questionnaire souvent utilisé pour évaluer de façon individuelle la QdV des patients (Hickey *et al*, 1996). Il s'agit d'un hétéro-questionnaire qui doit être rempli lors d'un entretien avec un psychologue ou un attaché de recherche clinique. On demande ainsi au patient de sélectionner les cinq domaines les plus importants pour sa QdV puis d'évaluer son niveau actuel dans chacun de ces domaines sur une échelle visuelle analogique. Un indice global est généré résumant la satisfaction globale de chaque domaine et tenant compte de leur importance relative.

3.6. Les questionnaires spécifiques du cancer

3.6.1. Le groupe Qualité de Vie de l'EORTC

- **Le questionnaire QLQ-C30**

Le groupe QdV de l'« European Organization of Research and Treatment of Cancer » (EORTC) (http://groups.eortc.be/qol/quality-life-group) a développé et validé un questionnaire de QdV appelé QLQ-C30 spécifique du cancer (Aaronson *et al*, 1993). Ce questionnaire contient 30 items et permet d'évaluer 15 dimensions de QdV (Annexe A) :

- cinq échelles fonctionnelles : physique, rôle, émotionnelle, cognitive et sociale,
- une dimension de QdV/santé globale,
- huit échelles symptomatiques : fatigue, nausée et vomissement, douleur, dyspnée insomnie, perte d'appétit, constipation et diarrhée,
- ainsi que les difficultés financières liées à la maladie.

Les 28 premiers items sont construit sur une échelle de Likert à 4 modalités de réponse de type : 1 « Pas du tout » / 2 « Un peu » / 3 « Assez » / 4 « Beaucoup ». Les deux derniers items sont construits sur une échelle à 7 modalités de réponse. Ces deux items évaluent respectivement l'état physique et la QdV globale du patient ; la modalité de réponse 1 correspondant à un état « très mauvais » et la modalité de réponse 7 à un « excellent » état.

Un manuel de scoring a été créé par l'EORTC afin de donner les principales recommandations de calcul des scores (Fayers PM *et al*, 2001). Un score est ainsi calculé par dimension et standardisé de 0 à 100 de sorte qu'un score faible corresponde à un faible niveau fonctionnel, un faible niveau de QdV/santé globale et à une absence de symptôme ; et un score élevé corresponde à un haut niveau fonctionnel, un haut niveau de QdV/santé globale et à une présence élevée de symptômes. Pour qu'un score puisse être estimé pour une dimension donnée, l'EORTC recommande qu'au moins 50% des items de la dimension considérée soient renseignés. Par défaut, la méthode d'imputation des items manquants proposée par l'EORTC est la méthode d'estimation simple par la moyenne, en considérant donc que l'item manquant ne diffère pas significativement des items renseignés. L'item manquant est alors remplacé par la moyenne des items répondus pour un individu donné.

- **Modules supplémentaires**

Selon les localisations cancéreuses et les modalités de prise en charge, différents modules supplémentaires ont été développés par l'EORTC. Ces modules sont à administrer conjointement avec le questionnaire core QLQ-C30. Ainsi, de nombreux modules supplémentaires ont été créés tels que les modules QLQ-BR23 pour le cancer du sein (Sprangers *et al*, 1996), QLQ-LC13 pour le cancer du poumon (Bergman *et al*, 1994) ou QLQ-PR25 pour le cancer de la prostate (van Andel *et al*, 2008). A titre d'illustration, le questionnaire QLQ-BR23 spécifique du cancer du sein (Annexe B) contient 23 items permettant d'évaluer 4 dimensions fonctionnelles (image corporelle, fonctionnement sexuel, plaisir sexuel et perspectives futures) et 4 dimensions symptomatiques (symptômes liés au traitement, symptômes au niveau du bras, symptôme au niveau du sein, inquiétude liée à la perte des cheveux) spécifiques du cancer du sein et de ses modalités de traitement (Sprangers *et al*, 1996). La méthode de scoring des modules supplémentaires est la même que celle du questionnaire core QLQ-C30.

Des modules ont également été développés pour une population spécifique comme le module QLQ-ELD14 spécifique des personnes âgées atteintes de cancer (Wheelwright *et al*, 2013). Enfin, une version réduite du questionnaire QLQ-C30 a été créée spécifiquement pour les patients en situation palliative. Il s'agit du questionnaire QLQ-C15PAL contenant ainsi 15 items du questionnaire QLQ-C30 (Groenvold *et al*, 2006). L'ensemble des modules validés de l'EORTC sont résumés en Annexe C.

De nouveaux modules sont également sans cesse en cours de développement comme le module QLQ-BrR24 spécifique de la reconstruction mammaire (Winters *et al*, 2014). Un module spécifique de la fatigue liée au cancer est également en cours de développement (Weis *et al*, 2013). De plus, les modules existants sont également régulièrement révisés, tant au niveau de la formulation des items eux-mêmes qu'au niveau du nombre d'items. Ainsi, le module QLQ-CR38 initialement développé pour le cancer colorectal (Sprangers *et al*, 1999a) en 1999 a été remplacé en 2009 par le module QLQ-CR29 plus court (38 items initialement versus 29 items dans la version actuelle) (Whistance *et al*, 2009).

Les questionnaires de QdV de l'EORTC sont développés d'emblée dans différentes langues et cultures. L'EORTC a développé une approche en quatre phases pour développer des modules de QdV (Blazeby *et al*, 2001; Sprangers *et al*, 1993). Ces quatre phases sont résumées dans le Tableau 1 ci-après.

Tableau 1 : Guidelines pour le développement d'un module de QdV de l'EORTC

Phase	Objectif	Procédure
1	Génération des domaines de QdV pertinents pour le groupe de patients sélectionnés	Revue de la littérature Entretiens semi-structurés avec des professionnels de santé et des patients Analyse de données qualitative et quantitative Combinaison des résultats issus des entretiens pour produire une liste de domaines/dimensions
2	Construction d'un questionnaire provisoire	Consultation de la base de données d'item du groupe QdV de l'EORTC Construction de nouveaux items Traduction du questionnaire provisoire
3	Test pilote du questionnaire concernant l'acceptabilité et la pertinence	Complétion du module par des patients lors d'un entretien Analyse de données quantitative et qualitative Modification du questionnaire Rapport officiel de développement examiné par le groupe QdV de l'EORTC
4	Test dans un contexte international	Test psychométrique de la validité, la fiabilité et la sensibilité du module

L'ensemble des modules de QdV et de PROs de l'EORTC en cours de développement sont résumés en Annexe D.

3.6.2. Le groupe FACT

- **Questionnaire FACT-G**

Le groupe américain « Functional Assessment of Cancer Therapy » (FACT) (http://www.facit.org/) a également développé et validé un questionnaire de QdV appelé FACT-General (FACT-G) spécifique du cancer (Cella *et al*, 1993). Ce questionnaire contient 27 items construits sur une échelle de Likert à 5 modalités de réponses codées : 0 « Pas du tout » / 1 « Un peu » /2 « Moyennement » / 3 « Beaucoup » / 4 « Enormément ». Quatre dimensions de bien-être sont évaluées par le biais de ces items : le bien-être physique, social/familial, émotionnel et fonctionnel.

Comme pour les questionnaires de l'EORTC, un score est calculé par dimension. Ces scores varient de 0 à 28 où 0 représente un bas niveau de bien-être et 28 un haut niveau de bien-être. Pour qu'un score puisse être estimé, au moins 50% des items de la dimension doivent être renseignés. Contrairement au questionnaire QLQ-C30, les items du questionnaire FACT-G sont séparés par dimension et le nom de chaque dimension évaluée précède les items correspondant. De plus, un score FACT-G global est estimé à partir des scores obtenus pour chaque dimension. Ce score correspond à la somme des scores bruts obtenus pour chaque dimension et varie de 0 (faible niveau) à 108 (bon niveau de bien-être général). Ce score global ne peut être calculé que si l'ensemble des sous-scores ont pu être estimés. Les quatre scores de bien-être ainsi que le score FACT-G global sont généralement standardisés de 0 à 100 dans le but de faciliter la comparaison avec d'autres questionnaires.

- **Dimension additionnelle spécifique d'une localisation cancéreuse**

Afin d'évaluer les symptômes et modalités de traitement propres à chaque localisation cancéreuse, le choix du groupe FACT a été d'ajouter une sous-échelle spécifique de la localisation concernée au sein du questionnaire FACT-G. A titre d'exemple, le questionnaire FACT-B spécifique du cancer du sein développé par Brady *et al.* contient les 27 items

mesurant les quatre dimensions de bien-être du FACT-G ainsi que 9 items supplémentaires permettant d'évaluer une sous-échelle spécifique du cancer du sein (Brady *et al*, 1997). Ainsi, un seul questionnaire est administré au patient. D'autre part, un score global FACT-B est estimé et correspond à la moyenne des sous-scores obtenus pour chaque dimension de bien-être et pour la sous-échelle additionnelle spécifique au cancer du sein. Enfin, selon les questionnaires, des scores indices sont également proposés. Concernant le questionnaire FACT-B, l'indice « Trial Outcome Index » (TOI) correspond à un score résumé du bien-être physique et fonctionnel et de la sous-échelle spécifique au cancer du sein. Ces scores résumés (score global FACT-G, score global FACT-L et indice TOI) peuvent alors être utilisés dans les analyses de données de QdV.

- **Dimension additionnelle spécifique d'un domaine**

Afin d'évaluer un domaine de QdV spécifique ou un symptôme particulier, le groupe FACT a également développé des questionnaires contenant les items du FACT-G ainsi qu'une dimension additionnelle mesurant un domaine particulier, telle que la fatigue. Ainsi, le FACT-F contient les 27 items du FACT-G ainsi que 13 items supplémentaires évaluant le niveau de fatigue du patient (Yellen *et al*, 1997).

Contrairement aux questionnaires de QdV du groupe EORTC, les questionnaires du groupe FACT sont dans un premier temps développés et validés en anglais selon la « culture nord-américaine ». Dans un second temps, ils sont adaptés et validés dans différentes langues. Les groupes EORTC et FACT présentent ainsi certaines différences de construction et de conceptualisation des questionnaires de QdV. Les principales différences sont résumées dans le Tableau 2 ci-après.

<u>Tableau 2</u> : Différence de conceptualisation des questionnaires de QdV entre les groupes EORTC et FACT

	EORTC	FACT
Domaines	Centrés sur les symptômes et dimensions fonctionnelles	Dimensions de bien-être et dimensions de préoccupations additionnelles
Structure	Items non séparés par domaine	Items séparés par domaine
	Domaines de QdV évalués non énoncés	Items précédés par une étiquette indiquant la dimension mesurée : ex: "bien-être physique"
Echelles	Uni et multi-items	Multi-items
Modalité de réponse par item	4 modalités	5 modalités
Questions formulées	à la seconde personne	à la première personne
	ex : « Avez-vous eu des nausées ? »	ex : « J'ai des nausées »
Par dimension	Items formulés dans le même sens	Nécessité d'inverser certains items pour le calcul du score
Score de QdV globale	Evalué par deux items du QLQ-C30	Moyenne des scores obtenus pour chaque dimension
Par localisation cancéreuse	1 module complémentaire	1 dimension additionnelle au questionnaire FACT-G
	Module à administré conjointement avec le QLQ-C30	1 seul questionnaire à administrer
	Plusieurs dimensions spécifiques de la localisation	1 seule dimension spécifique de la localisation
Développement	Simultanément dans plusieurs langues et cultures	Dans une langue puis de façon séquentielle adaptation à d'autres langues et cultures

3.6.3. <u>Exemples d'autres questionnaires</u>

Les groupes EORTC et FACT sont les deux grands groupes développant et validant des auto-questionnaires de QdV spécifiques du cancer. Il existe cependant d'autres auto-questionnaires spécifiques du cancer.

- **Rotterdam Symptom Checklist**

Le Rotterdam Symptom Checklist est également un questionnaire de QdV spécifique du cancer (de Haes *et al*, 1990). Ce questionnaire est assez proche dans sa structure du questionnaire EORTC QLQ-C30, à la fois au niveau du nombre de questions et des domaines abordés. Il contient ainsi 30 items construits sur une échelle de Likert à 4 modalités de réponse (« Pas du tout » / « Un peu » / « Assez / « Beaucoup »), une question concernant le niveau d'activité et une question sur une échelle de 1 à 7 concernant le niveau de QdV

globale. Ce questionnaire est centré sur les symptômes et les effets secondaires les plus couramment exprimés par les patients atteints d'un cancer. Ce questionnaire était souvent utilisé dans les études européennes dans le passé. Du fait de l'existence du groupe EORTC, ce questionnaire est peu utilisé à ce jour.

- **Functional Living Index - Cancer**

Le questionnaire « Functional Living Index – Cancer » (FLIC) est également un questionnaire multidimensionnel spécifique du cancer (Schipper *et al*, 1984). Il contient 22 items et permet d'évaluer la qualité fonctionnelle de la vie quotidienne de patients atteints d'un cancer en évaluant 5 dimensions : la QdV globale, le rôle fonctionnel, l'état émotionnel, la sociabilité, les douleurs et nausées. Un score est calculé par dimension et standardisé de 0 à 100 de sorte qu'un score élevé représente un bon niveau de QdV (niveau fonctionnel élevé et peu de symptômes). Ce questionnaire a été validé en langue française (Mercier *et al*, 1998). Il est majoritairement utilisé dans les études nord-américaines.

- **Expanded Prostate Index Cancer**

Le questionnaire « Expanded Prostate Index Cancer » (EPIC) est quant à lui spécifique du cancer de la prostate (Wei *et al*, 2000). Ce questionnaire conteint 50 items répartis en 4 domaines (urinaire, digestif, sexuel et hormonal) et 8 sous-échelles (fonctionnelles et symptomatiques) ainsi qu'une question de satisfaction globale. Les items sont séparés par dimension et les domaines évalués sont clairement énoncés dans le questionnaire. Chaque item possède 4 ou 5 modalités de réponse. Un score est calculé pour chaque sous-échelle et chaque domaine (Wei, 2002). Des items sont formulés positivement et d'autres négativement ce qui nécessite une inversion du codage de certains items lors du calcul du score. Pour ce questionnaire, un score élevé représente un haut niveau de QdV. Ainsi, un score élevé pour une sous-échelle fonctionnelle représente un niveau fonctionnel élevé ; un score élevé pour une sous-échelle symptomatique représente un bas niveau symptomatique. Le pourcentage de données manquantes toléré pour pourvoir estimer un score est de l'ordre de 20%. Ce questionnaire a été développé et validé en langue anglaise (Wei *et al*, 2000) et est en cours de validation en langue française (Mariet *et al*, 2014).

- **Linear Analogue Self-Assessment**

Le questionnaire « Linear Analogue Self-Assessment » (LASA) est un instrument spécifique du cancer où chaque domaine est évalué par une échelle visuelle analogique (Selby et al, 1984). Ce questionnaire contient une échelle évaluant la QdV globale et 30 échelles évaluant une dimension spécifique de la QdV: 18 échelles de santé générale et 12 échelles liées au traitement ou à la maladie. Il est demandé aux patients d'évaluer dans quelle mesure chacun des aspects de leur vie énoncés est affecté par leur état de santé du jour. Chaque échelle mesure 10 cm de long et chaque extrémité de l'échelle est labélisée d'une étiquette. Par exemple, pour une échelle mesurant les nausées, les extrémités de l'échelle sont labélisées respectivement « nausée extrêmement élevée » et « pas de nausée ». Un score est mesuré de 0 à 10 pour chaque échelle par lecture de la valeur rapportée par le patient tel qu'un score élevé corresponde à un niveau de QdV élevé ou à une absence de symptômes.

- **Le questionnaire de Spitzer**

Le Spitzer QoL Index est un questionnaire générique contenant 5 items et une échelle visuelle analogique évaluant la QdV globale (Spitzer *et al*, 1981). Les items évaluent l'activité, la vie quotidienne, la santé, le soutien de la famille et des amis, ainsi que l'attitude générale du patient. Chaque item est construit sur une échelle de Likert à 3 modalités de réponse. Un score global de 0 à 10 est calculé d'après ces items tel que 10 représente le meilleur niveau de QdV possible. Deux versions de ce questionnaire existent : une version auto-évaluative et une version hétéro-évaluative. Le questionnaire Spitzer peut donc être rempli par une tierce personne et donner ainsi une mesure proxy de la QdV du patient.

Des hétéro-questionnaires spécifiques du cancer ont également été développés pour être énoncés au patient par une tierce personne. A titre d'illustration, le questionnaire QUAL-E (« QUAlity of Life at the End of life ») spécifique du cancer pour les patients en fin de vie (Steinhauser *et al*, 2004) a été conçu pour être énoncé au patient par un attaché de recherche clinique, compte tenu de l'état de fatigue et de l'incapacité du patient en situation palliative à remplir le questionnaire lui-même. Ce questionnaire contient 26 items construits sur une échelle de Likert à 5 modalités de réponses. Il est également demandé aux patients de citer

trois symptômes particulièrement ressentis durant les 4 dernières semaines. Cinq dimensions de QdV sont évaluées : la sévérité des symptômes, les relations avec le système de soin, le soutien affectif social, la préparation à la fin de la vie et le sentiment de complétion de la vie.

4. Les propriétés psychométriques des questionnaires et méthodes de validation

Les questionnaires de QdV doivent respecter un certain nombre de propriétés psychométriques attestant de leur robustesse et de leur fiabilité (Fayers & Machin, 2000). Ces propriétés doivent donc être étudiées et validées lors de la conception du questionnaire. De plus, lorsqu'un questionnaire est adapté dans une autre langue que celle dans laquelle il a été développé, cela requiert une adaptation transculturelle puisqu'une simple traduction ne peut être suffisante (Guillemin *et al*, 1993). Une validation des propriétés psychométriques de la version traduite du questionnaire doit donc à nouveau être réalisée. Enfin, lorsqu'une version réduite d'un questionnaire existant est proposée, les propriétés psychométriques doivent également être étudiées afin de s'assurer que les informations retenues par les items sélectionnés sont suffisantes et bien représentatives de l'ensemble des dimensions du questionnaire complet.

Ces propriétés sont le plus souvent étudiées selon la théorie classiques des tests (« Classical Test Theory », CTT) considérant que le score calculé d'après les réponses aux items est une bonne représentation du niveau de QdV du patient. Cependant, depuis une décennie, la théorie moderne de réponse à l'item (« Item Response Theory », IRT) est également utilisée pour la validation des questionnaires de QdV. Ces deux approches sont complémentaires et une utilisation conjointe parait indispensable pour s'assurer de la robustesse de l'instrument de mesure.

4.1. Propriétés psychométriques et méthodes classiques de validation

4.1.1. La validité

La validité d'un questionnaire correspond à la capacité de ce questionnaire à mesurer le (ou les) domaine(s) de QdV qu'il est censé mesurer.

- La **validité apparente** (« face validity ») est une première mesure, davantage qualitative, de la validité du questionnaire. Elle permet de s'assurer que les items couvrent bien a priori les domaines qu'ils sont censés mesurer et sans ambiguïté. La « face validity » est généralement mesurée par le biais d'une évaluation par les patients et les experts médicaux de la validité du questionnaire. Elle est souvent évaluée auprès d'un sous-échantillon de patients au moment du pré-test du questionnaire. On recueille alors l'avis du patient vis-à-vis du questionnaire, soit lors d'un entretien avec un attaché de recherche clinique par exemple, soit par le biais d'un questionnaire de débriefing auto-administré. Cette évaluation porte principalement sur la compréhension et la formulation des questions, leur redondance et leur pertinence. On demande par exemple aux patients si le questionnaire leur a semblé trop long ou trop compliqué, si certaines questions sont dérangeantes, confuses, redondantes ou au contraire si certains aspects importants de leur QdV, certains effets du traitement ou symptômes de la maladie n'ont pas été abordés.

La « face validity » correspond également à une mesure de l'acceptabilité du questionnaire par le patient : pour qu'un questionnaire puisse être utilisé, il doit être accepté par le patient, i.e. que ce dernier doit accepter de remplir le questionnaire proposé. Les questions doivent donc être compréhensibles et ne doivent être ni confuses ni dérangeantes pour le patient. Un questionnaire trop long, trop compliqué ou redondant pourrait ainsi lasser le patient, ce qui le conduirait à ne pas remplir l'intégralité du questionnaire.

L'acceptabilité d'un questionnaire est ainsi mesurée par :

- Le taux de participation des patients ;
- le taux de complétion des questionnaires retournés : les questionnaires doivent présenter peu de questions ou de dimensions manquantes ;
- et le temps de remplissage du questionnaire.

Des taux élevés de participation et de complétion des questionnaires permettront de s'assurer de l'absence d'un biais de sélection de populations. En effet, un mauvais taux de complétion pourrait conduire à mener l'analyse sur une sous-population de patients en meilleur état de

santé et/ou une population particulière telle que des patients plus jeunes que la population initialement ciblée.

L'acceptabilité d'un questionnaire est également illustrée par l'absence d'effet plafond et/ou d'effet plancher. L'effet plafond est illustré par un nombre élevé de patients ayant obtenu le plus haut score qu'un patient puisse obtenir pour une dimension donnée en déclarant par exemple un niveau de symptôme maximal ou un niveau fonctionnel maximal. Inversement, l'effet plancher correspond à un taux élevé de patients ayant obtenu le score le plus bas qu'un patient puisse obtenir pour une dimension donnée. L'existence d'effet plafond et/ou d'effet plancher pour une dimension donnée sous-entendrait que les modalités de réponse aux items ne sont pas suffisantes ou bien que les items eux-mêmes ne sont pas adaptés à la population d'étude. Le pourcentage de patients ayant obtenu les scores les plus faibles et les plus élevés possibles pour une dimension donnée doivent donc être limités (sans pour autant être nul).

- La **validité de contenu** (« content validity ») concerne l'adéquation du contenu d'un instrument en termes de nombre d'items et de leur champ/portée.

La validité de contenu est assez proche de la validité apparente et ces deux aspects peuvent être confondus. La différence entre les deux types de validité est que la validité apparente correspond à une revue critique des items après le développement du questionnaire, alors que la validité de contenu est réalisée au cours du développement du questionnaire et consiste à s'assurer que celui-ci couvre toutes les questions importantes.

- La **validité de structure** (« construct validity ») est une des caractéristiques les plus importantes de l'instrument de mesure. Elle correspond à une évaluation de la mesure dans laquelle le questionnaire mesure bien le concept qu'il est censé mesurer. Elle permet de vérifier :
    - La dimensionnalité du questionnaire : est-ce que tous les items d'une même sous-échelle sont liés à une seule variable latente ou bien davantage de variables latentes sont-elles nécessaires pour expliquer la variabilité observée ?
    - L'homogénéité des items: est-ce que tous les items d'une même sous-échelle sont bien fortement liés à cette échelle ?
    - Le chevauchement éventuel entre les différents domaines : est-ce que certains items d'une sous-échelle donnée sont également liés à une autre échelle du questionnaire ?

La validité de structure se décompose généralement en trois types de validité : la validité convergente, la validité divergente et la validité de groupes connus.

Les **validités convergente et divergente** sont généralement évaluées par le biais de la matrice de corrélation « multi-traits multi-méthodes ». Ainsi, les items d'une même sous-échelle sont censés être davantage corrélés aux autres items de la même sous-échelle qu'aux autres domaines du questionnaire (matrice multi-traits). Si les items sont également comparés aux domaines d'un autre questionnaire, on parle de matrice multi-méthodes. La validité convergente est évaluée à l'aide du coefficient de corrélation de Spearman entre l'item et le score obtenu à sa propre échelle, calculé en excluant l'item considéré. Concernant la validité divergente, la corrélation entre chaque item et le score obtenu à sa propre échelle est censée être plus élevée que la corrélation entre l'item et les scores obtenus pour les autres échelles du questionnaire.

Les analyses factorielles de type Analyses Factorielles des Correspondances Multiples (AFCM) ou Analyses en Composantes Principales (ACP) (Floyd & Widaman, 1995) permettent également de valider la structure interne du questionnaire, i.e. de vérifier que les items d'une même dimension sont bien davantage liés entre eux qu'aux items des autres dimensions. Les objectifs de ces deux analyses sont les mêmes, cependant les AFCM sont adaptées aux mesures catégorielles tandis que les ACP s'appliquent aux données continues. Un des objectifs de ces deux analyses est de représenter graphiquement et le plus fidèlement possible les liaisons entre les n variables initiales dans un sous-espace de dimension p tel que $p < n$. Ainsi, en appliquant ce type d'analyses aux n items d'un questionnaire permettant a priori d'évaluer p dimensions de la QdV, l'objectif et de retrouver ces p dimensions. Les items permettant d'évaluer une même dimension doivent être fortement corrélés à un certain axe i de l'analyse factorielle. Une rotation de type varimax est couramment appliquée afin de faciliter la lecture des axes.

Le principe de l'ACP est de construire de nouvelles variables synthétiques qui sont des combinaisons linéaires des variables initiales, de plus grande variance possible. Ces nouvelles variables (appelées composantes principales) captureront la plus grande part de la variabilité des données. Ainsi, une première variable C_1 est construite telle que la variance de C_1 est maximale. Puis, on itère la procédure en construisant une seconde variable C_2 telle que C_2 n'est pas corrélée avec C_1 et est de variance maximale. En général, on s'intéresse aux variations par rapport à l'individu moyen. Les variables d'origines sont donc centrées et nécessairement, les composantes principales sont aussi de moyenne nulle. Les solutions sont

alors les vecteurs propres associées aux plus grandes valeurs propres de la matrice de variance covariance des variables d'origines (ou variables centrées). Ainsi, la première composante principale C_1 est obtenue par combinaisons linéaires des n variables d'origines $X_1, \dots, X_n$ avec des poids donnés par les éléments du premier vecteur propre α_1 (associé à la plus grande valeur propre λ_1), tel que :

$$C_1 = \alpha_{11}X_1 + \cdots + \alpha_{1n}X_n$$
$$Var(C_1) = \lambda_1$$

De la même façon :

$$C_2 = \alpha_{21}X_1 + \cdots + \alpha_{2n}X_n$$
$$Var(C_2) = \lambda_2$$

Et les deux composantes ne sont pas corrélées, i.e. : $Cov(C_1, C_2) = 0$.

La **validité de groupes connus** (« known-group validity ») correspond à la capacité du questionnaire à mettre en évidence une différence entre deux groupes de patients ne présentant pas le même état de santé. Par exemple, on pourrait s'attendre à ce que des patients avec un cancer à un stade avancé aient un niveau de QdV plus bas que ceux ayant un cancer à un stade précoce. Un instrument de QdV valide est censé mettre en évidence une différence de niveau de QdV entre ces deux groupes.

- La **validité de critère** (« criterion validity ») consiste à comparer l'instrument en cours de développement à un instrument de référence (déjà validé) qui est censé mesurer le même concept ou les mêmes dimensions de QdV. Une matrice de corrélation entre les scores des différents questionnaires permet ainsi d'apprécier l'adéquation des dimensions censées mesurer le même domaine de QdV.

4.1.2. La fiabilité

La fiabilité correspond à la capacité d'un instrument de mesure à donner des résultats reproductibles et cohérents.

La **cohérence interne** des questionnaires est généralement appréciée à l'aide du coefficient alpha de Cronbach (Cronbach, 1951). Ce coefficient varie entre 0 et 1. Les items d'une dimension donnée présentent une bonne cohérence interne si le coefficient alpha de Cronbach

est au moins égal à 0.70 (Nunnally & Bernstein, 1994). La formule correspondante, se basant sur la variance totale et la variance due à chaque item, est la suivante :

$$\propto = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} Var(Y_i)}{Var(X)}\right)$$

où k est le nombre d'items de la dimension considéré, X le score total (somme des réponses aux items) et Y_i la réponse à l'item i. Ce coefficient ne peut être calculé que pour les échelles multi-items.

Ce coefficient peut alors être recalculé en excluant un item de la dimension. Si la cohérence augmente en excluant cet item, alors il pourrait être envisagé de supprimer l'item de la dimension.

Les questionnaires doivent également être **reproductibles**. Un patient resté dans un état de santé stable entre deux passations d'un même questionnaire doit donc fournir des réponses identiques lors des deux mesures. La reproductibilité (ou répétabilité) du questionnaire est mesurée selon la méthode « test-retest ». Cette méthode nécessite que les patients complètent deux fois le questionnaire. L'intervalle de temps entre les deux mesures doit être relativement court (environ une à deux semaines maximum) et le patient doit être dans un état de santé stable entre les deux passations. L'intervalle de temps entre les deux passations ne doit pas non plus être trop court puisque le patient pourrait se souvenir des réponses données lors de la première passation. La reproductibilité est alors généralement mesurée par le coefficient de corrélation intra-classe (CCI). Le CCI mesure la force de l'accord entre des mesures répétées. Il s'applique à des données continues, mais peut également être appliqué à des données catégorielles ordinales avec au moins 4 modalités de réponse. Il mesure la proportion de la variance totale qui est associée avec la variabilité inter-patient. La formule du coefficient CCI est la suivante :

$$CCI = \frac{\sigma^2_{patient}}{\sigma^2_{patient} + \sigma^2_{erreur}}$$

où $\sigma^2_{patient}$ est la variance inter-patient et σ^2_{erreur} la variance de l'erreur aléatoire. Ce coefficient peut être obtenu par une ANOVA. Si le coefficient est proche de 1, cela signifie

que la variance du terme de l'erreur aléatoire est faible. Ainsi, une grande proportion de la variance observée est due à la variance entre les sujets.

Un coefficient supérieur à 0.90 correspondra à une très bonne reproductibilité de l'échelle et supérieur à 0.70 à une bonne reproductibilité (Nunnally & Bernstein, 1994).

4.1.3. La sensibilité au changement

La sensibilité au changement est la capacité d'un questionnaire à déceler un changement de l'état de santé du patient sans qu'il soit nécessairement cliniquement pertinent pour le patient.

Différentes méthodes statistiques existent pour déterminer la sensibilité au changement d'un questionnaire. Il n'existe pas de consensus sur les méthodes statistiques à utiliser. Néanmoins, les deux indicateurs les plus couramment utilisés sont l'effet taille (« Effect Size ») (Kazis *et al*, 1989) et la réponse moyenne standardisée (« Standardized Response Mean ») (Katz *et al*, 1992).

L' « Effect Size » (ES) correspond au changement moyen entre les deux temps de mesure x_1 et x_2 divisé par l'écart type (SD) du premier temps de mesure :

$$ES = \frac{\overline{x_2} - \overline{x_1}}{SD(x_1)}$$

où $\overline{x_1}$ et $\overline{x_2}$ correspondent respectivement aux moyennes obtenus aux deux temps de mesure x_1 et x_2.

Le coefficient « Standardized Response Mean » correspond au changement moyen entre les deux temps de mesure divisé par l'écart type de la différence :

$$SRM = \frac{\overline{x_2} - \overline{x_1}}{SD(x_2 - x_1)}$$

Pour ces indicateurs, plus le coefficient est élevé, plus le changement est important (Cohen, 2013): un coefficient inférieur 0.2 en valeur absolue est considéré comme un changement non significatif, entre 0.2 et 0.5 comme modéré et supérieur à 0.5 comme large.

4.1.4. La « responsiveness »

La « responsiveness » est la capacité d'un questionnaire à déceler un changement de l'état de santé du patient qui soit cliniquement pertinent. Cette propriété relie donc la sensibilité au changement à un changement clinique pour le patient.

L'interprétation des scores de QdV et d'une différence de scores cliniquement pertinente entre deux temps de mesures est un problème majeur en QdV (Liang, 2000; Lydick & Epstein, 1993; Osoba *et al*, 1998; Stucki *et al*, 1996). Il existe une différence entre la notion de résultat statistiquement significatif et celle de résultat cliniquement pertinent. Un résultat statistiquement significatif peut ne pas être cliniquement pertinent pour le patient. En effet, plus la taille de l'échantillon considéré est grande, plus une différence de score entre deux temps de mesure, aussi petite soit-elle, peut être significative d'un point de vue statistique. Or, cette petite différence peut ne pas être cliniquement significative du point de vue du patient.

La différence minimale cliniquement importante (DMCI) a ainsi été définie comme la plus petite différence dans un score de QdV qui serait perçue comme ayant un sens clinique pour le patient (Jaeschke *et al*, 1989). Déterminer la DMCI pour l'interprétation des scores de QdV est indispensable tant pour pouvoir évaluer l'efficacité d'une prise en charge que pour déterminer le nombre de sujets nécessaires pour la réalisation d'un essai clinique.

Il n'existe pas à ce jour de méthode standard pour déterminer la DMCI (Terwee *et al*, 2010). Les méthodes proposées sont généralement regroupées en deux catégories : les méthodes basées sur l'ancre (« anchor-based method ») et les méthodes basées sur la distribution (« distribution-based method »).

Les méthodes basées sur l'ancre se focalisent sur la relation entre les scores de QdV étudiés et un critère externe ayant un sens clinique. Ce critère externe correspond à l'ancre. Cette ancre peut être un indicateur de la progression de la maladie ou de la réponse clinique, mais également une évaluation effectuée par le patient lui-même du changement de la QdV globale. Ce dernier type de critère externe est celui le plus souvent utilisé comme ancre pour déterminer la DMCI (Crosby *et al*, 2003; Revicki *et al*, 2008). A titre d'exemple, la question de la transition de Jaeschke est régulièrement utilisée comme ancre (Jaeschke *et al*, 1989). Cette question est généralement formulée de la façon suivante :

« Durant les trois derniers mois, considérez-vous que votre niveau de qualité de vie :

- N'a pas changé globalement

- S'est détérioré : un peu, assez, beaucoup
- S'est amélioré : un peu, assez, beaucoup ».

Plusieurs ancres peuvent également être utilisées dans la même étude afin de tester la cohérence entre les résultats (Yost *et al*, 2005).

Les méthodes basées sur la distribution considèrent que la DMCI peut être estimée en se basant sur la distribution des scores observés sur un échantillon représentatif de la population considérée. Certains chercheurs suggèrent que la moitié de l'écart type (Norman *et al*, 2003) ou l'erreur standard de mesure (« Standardized Response Mean ») (Wyrwich *et al*, 1999a; Wyrwich *et al*, 1999b) peuvent être considérés comme la DMCI. Cependant, cette différence observée est surement cliniquement importante mais ne correspond pas nécessairement à la différence minimale cliniquement importante. Il est nécessaire de résumer ces informations en termes d'« Effect Size » pour pouvoir apprécier le degré de changement. De plus, ces méthodes basées sur la distribution ne fournissent pas d'information directe sur la DMCI, seules les méthodes basées sur l'ancre permettent de mesurer directement la DMCI.

Pour les questionnaires de QdV de l'EORTC, une différence de 5 ou 10 points pour un score est généralement considérée comme la DMCI (Cocks *et al*, 2011; Osoba *et al*, 1998). En ce qui concerne les questionnaires du groupe FACT, une différence d'au moins 5 à 7 points pour le score global est généralement considérée comme la DMCI (Cella *et al*, 1993).

Cependant, comme la DMCI dépend des caractéristiques de la population et du contexte clinique de l'étude, il n'existe pas nécessairement une seule valeur de DMCI pour un même questionnaire pour toutes les populations considérées (Revicki *et al*, 2008). De plus, la DMCI pour un même questionnaire peut dépendre du sens du changement considéré, i.e. peut varier selon que l'on considère une amélioration du score ou une détérioration (Cella *et al*, 2002). Généralement, elle est plus importante pour la détérioration que pour l'amélioration (Hong *et al*, 2013). Une autre problématique de la DMCI concerne l'influence du niveau de QdV des patients à l'inclusion (Ward *et al*, 2014). Certaines études ont ainsi proposé de définir la DMCI en fonction du niveau du score à l'inclusion et non comme une valeur unique (Crosby *et al*, 2003; Revicki *et al*, 2008). En effet, une valeur de DMCI unique suppose que le score a des propriétés de mesures d'intervalles, i.e. qu'une augmentation ou une diminution de Δ points du score du patient a la même signification au regard du vrai niveau de QdV du patient quel que soit le niveau du score initialement observé. Les scores obtenus d'après les réponses

aux items respectent rarement cette propriété. Ces scores correspondent davantage à des mesures ordinales, ayant des propriétés d'ordonnancement, i.e. qu'on peut classer les patients selon les valeurs de scores observées : par exemple, un patient avec un score faible présente un moins bon niveau de QdV qu'un patient avec un score élevé. Ainsi, une différence de Δ points sur une échelle de 0 à 100 n'aura peut-être pas nécessairement la même signification pour un patient avec un score faible correspondant à un bas niveau de QdV que pour un patient avec un score élevé et correspondant à un haut niveau de QdV. Enfin, la problématique des effets planchers et des effets plafonds compromet également l'observation d'un changement important même si ce changement a lieu.

Les DMCI retenues généralement pour les questionnaires EORTC et FACT sont donc très approximatives et susceptibles de varier entre les populations et les questionnaires utilisés. Cependant, ces valeurs sont généralement celles utilisées dans les études longitudinales pour une question de simplicité.

Les différentes méthodes présentées ci-dessus pour la validation des propriétés psychométriques correspondent à la théorie classique des tests.

4.2. Approche moderne de la théorie de réponse à l'item

Depuis une décennie, des modèles issus de la théorie de réponse à l'item (« Item Response Theory », IRT) sont également utilisés pour valider un questionnaire en cours de développement (Edelen & Reeve, 2007), tester la validité d'un questionnaire déjà existant et validé par la théorie classique (Jafari *et al*, 2012), réduire un questionnaire (Petersen *et al*, 2006) ou tester la présence d'un fonctionnement différentiel de l'item (Hardouin *et al*, 2012).

L'approche classique (CTT) et l'approche IRT sont complémentaires. La validation des propriétés psychométriques des questionnaires par les modèles IRT parait désormais indispensable compte tenu de leur bonne adaptation aux données issues des questionnaires. De plus, ces modèles ont des propriétés intéressantes telles que l'objectivité spécifique et leur robustesse face aux données manquantes (De Ayala, 2009).

Parmi les modèles IRT, deux grandes familles se distinguent : il s'agit des modèles de la famille de Rasch et ceux de la famille de Lord. Ces deux familles appréhendent les données

de manières opposées : pour les modèles de la famille de Rasch, ce sont les données qui doivent adhérer et s'adapter aux modèles ; en revanche, pour les modèles de la famille de Lord, c'est le modèle qui doit s'adapter le plus possible aux données.

Dans la théorie des tests classiques (CTT), le score obtenu d'après les réponses aux items est considéré comme une bonne représentation du trait latent. Les analyses sont alors réalisées d'après ces scores. Dans les modèles IRT, les items jouent un rôle clé. Ces modèles cherchent à mesurer directement le trait latent, soit la QdV. Le score brut n'est utilisé que comme une mesure ordinale permettant d'ordonner les individus. L'objectif des modèles IRT est de représenter sur un même continuum latent les items et les individus.

La majorité des modèles IRT reposent sur trois hypothèses fondamentales :

- L'**unidimensionnalité** du trait latent : tous les items doivent mesurer une seule dimension de QdV ;
- L'**indépendance locale** : les réponses aux items doivent être indépendantes les unes des autres conditionnellement à la valeur du trait latent. Autrement dit, la réponse à un item ne doit pas dépendre des réponses aux items précédents. A titre d'exemple, deux items imbriqués (où la réponse à l'item 2 ne peut être faite que si l'individu a répondu positivement à l'item 1) ne peuvent pas suivre un modèle de type IRT. ;
- La **monotonicité** : la probabilité d'une réponse positive (ou au moins celle-ci) à un item augmente avec la valeur de la variable latente.

4.2.1. Les modèles de la famille de Rasch

- **Le modèle de Rasch**

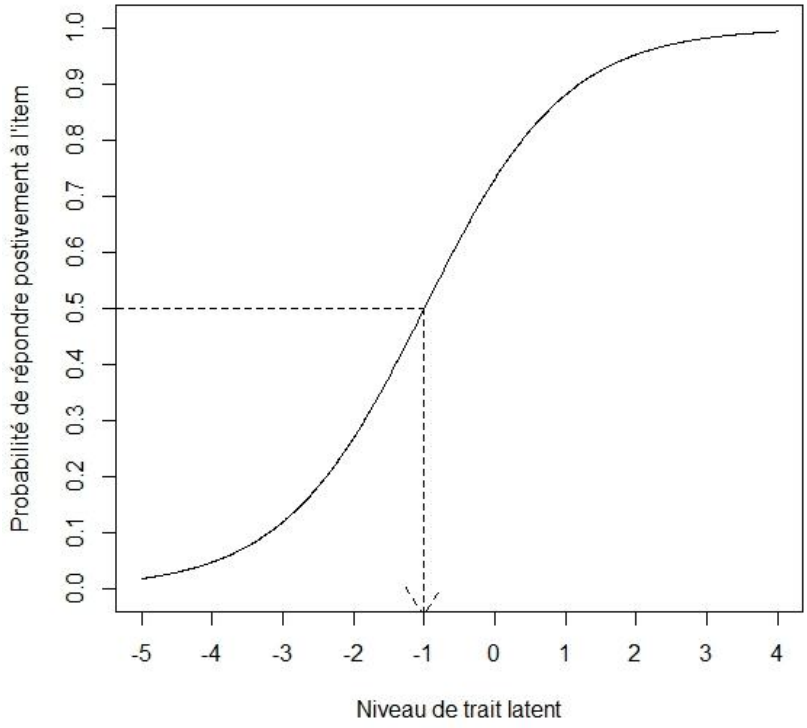
Le modèle de Rasch est un modèle pour item dichotomique (Rasch, 1993), i.e. dont les réponses sont de type « Oui » vs. « Non », codées 0 vs. 1. Ce modèle relie les items au trait latent par un lien logistique. Ils introduisent la notion de la **difficulté de l'item**. Pour un item dichotomique (0/1), la difficulté de l'item correspond à la probabilité de répondre positivement à cet item, i.e. de choisir la modalité 1. La difficulté de l'item représente le niveau de trait qu'il faut atteindre pour avoir une chance sur deux de répondre positivement à l'item. L'individu est caractérisé par sa capacité soit son niveau de trait latent.

Dans le modèle de Rasch, la probabilité p_{ij} qu'a l'individu i de répondre positivement à l'item j connaissant son niveau de trait latent θ_i et la difficulté δ_j de l'item j est donnée par la formule suivante:

$$p_{ij} = P(X_{ij} = 1|\theta_i, \delta_j) = \frac{\exp(\theta_i - \delta_j)}{1 + \exp(\theta_i - \delta_j)}$$

Pour pouvoir apprécier visuellement la bonne construction d'un item, on représente généralement les courbes caractéristiques des items (« Item Charateristic Curve », ICC). Pour un item dichotomique, ces courbes représentent la probabilité de choisir la réponse positive à un item (réponse 1) selon la valeur du trait latent.

A titre d'illustration, la Figure 3 ci-après représente la courbe caractéristique d'un item dichotomique de coefficient de difficulté δ = -1 selon un modèle de Rasch. Le niveau de difficulté de l'item peut se retrouver par lecture graphique : il s'agit du niveau de trait latent correspondant à une probabilité 0.5 de répondre positivement à l'item (correspondant à la flèche en pointillé).



<u>Figure 3</u> : Courbe Caractéristique d'un item de niveau de difficulté δ = -1 selon le modèle de Rash

Un item de coefficient de difficulté inférieur à -1 sera dit facile ; entre -1 et 1 la difficulté est considérée comme moyenne ; tandis qu'un item de coefficient de difficulté supérieur à 1 sera dit difficile.
Pour une échelle constituée de plusieurs items dichotomiques construis selon un modèle de Rasch, les courbes ICC ne se croisent pas puisque la pente est identique pour tous les items.

- **Propriétés des modèles de Rasch**

Les modèles de Rasch ont des propriétés psychométriques intéressantes telles que l'exhaustivité du score et l'objectivité spécifique (De Ayala, 2009).

L'exhaustivité du score brut observé (la somme des réponses aux items) implique que toute l'information nécessaire pour déterminer la valeur du trait latent est contenue dans le score. Ainsi, quel que soit la façon d'obtenir le score (donc quel que soit les réponses aux items), pour une même valeur de score donnée, deux individus n'ayant pas obtenu les mêmes réponses aux items auront le même score et donc la même valeur de trait latent. A titre d'illustration, si deux individus ont répondu positivement à 2 items parmi 4, ils auront le même score, et ce même si le premier individu a répondu positivement aux items les plus faciles (i.e. de coefficients de difficulté les plus faibles) tandis que le second a répondu positivement aux items les plus difficiles.

L'objectivé spécifique induit que les niveaux de difficulté des items sont indépendantes de la distribution du trait latent. De la même façon, les capacités (ou niveau de QdV) des individus sont indépendantes des difficultés des items. En conséquence, si l'échelle étudiée vérifie un modèle de Rasch, alors les niveaux de QdV des individus seront les mêmes quel que soit les items « retenus » pour évaluer leur niveau de QdV.

- **Modèles de Rasch pour items polytomiques**

Le modèle de Rasch ne peut s'appliquer qu'aux items dichotomiques. Or, ce type d'items est rarement utilisé dans les questionnaires de QdV. Des extensions de ce modèle dans le cadre polytomique ont donc été proposées (Masters, 1982; RJ, 2009). Dans la famille de Rasch, il existe ainsi deux modèles d'IRT adaptés aux items polytomiques. Il s'agit du modèle à crédit partiel (« Partial Credit Model », PCM) (Masters, 1982) et du modèle à échelle de classement

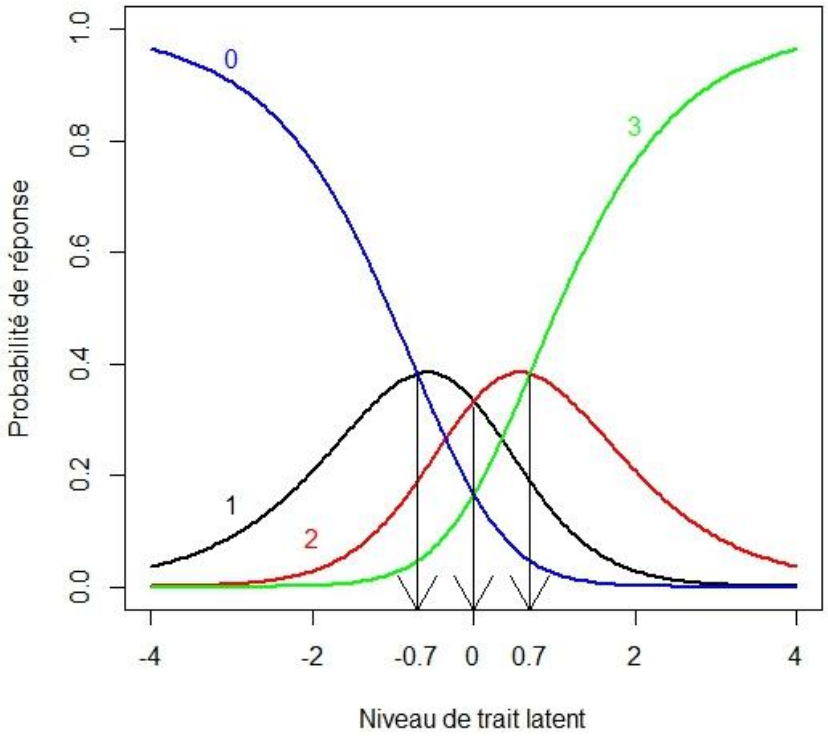
(« Rating Scale Model », (RSM)) (Andrich, 1978). Le modèle RSM est un sous-modèle du modèle PCM.

L'idée proposée par Masters pour développer le modèle PCM pour items polytomiques est de décomposer les réponses en une série de paires ordonnées de modalités de réponses adjacentes (Masters, 1982). Un modèle de Rasch pour item dichotomique est ensuite appliqué de façon successive à chaque paire Ainsi, un item à m_j modalités de réponse est vu comme une succession d'étapes à franchir pour pouvoir obtenir un score élevé à cet item (i.e. choisir une modalité élevée) où un « crédit partiel » est accordé à chaque nouvelle étape franchie avec succès.

Dans le modèle PCM, il existe autant de paramètre de difficulté que de (modalité – 1). Par exemple, pour un item polytomique à quatre modalités de réponses (codées 0/1/2/3), 3 paramètres de difficulté δ_k sont estimés, avec k ϵ (1, 2, 3). Le paramètre de difficulté δ_k représente le niveau de trait latent qu'il faut atteindre pour avoir une chance sur deux de choisir la modalité (k-1) ou la modalité k à l'item donné. Dans ce modèle, le nombre de modalités de réponse peut varier d'un item à un autre. La probabilité p_{ijk} qu'a l'individu i de choisir la modalité de réponse k à l'item j connaissant son niveau de trait latent θ_i et les niveaux de difficultés $\delta_{j,l}$ de l'item j, avec l=1, …, m_j, et m_j modalités de réponse est :

$$p_{ijk} = P\left(X_{ij} = k \middle| \theta_i,\ \delta_{j,1}, \dots, \delta_{j,m_j}\right) = \frac{\exp(k\theta_i - \sum_{l=1}^{k} \delta_{j,l})}{1 + \sum_{h=1}^{m_j} \exp(h\theta_i - \sum_{l=1}^{h} \delta_{j,l})}$$

A titre d'illustration, la Figure 4 représente la courbe caractéristique d'un item polytomique à 4 modalités de réponse (0, 1, 2, 3) selon un modèle PCM. Si le trait latent correspond à une échelle de QdV, alors un individu avec un niveau de QdV θ inférieur à -0.7 sera plus susceptible de choisir la modalité 0 à l'item. Si son niveau de QdV est relativement moyen (θ = 0), alors il aura autant de chance de choisir la modalité 1 ou la modalité 2. Pour un individu avec un niveau de QdV élevé (θ > 0.7), il est davantage probable qu'il choisisse la modalité de réponse 3.



<u>Figure 4</u> : Courbe caractéristique d'un item polytomique à 4 modalités de réponse selon un modèle PCM de difficulté de modalités de réponse δ_1=-0.7 ; δ_2=0 et δ_3=0.7

Le modèle RSM est un cas particulier du modèle PCM dans lequel tous les items ont le même nombre de modalités de réponse (Andrich, 1978). Dans ce modèle, un seul paramètre de difficulté par item est estimé. Ce paramètre correspond à la difficulté pour passer de la modalité 0 à la modalité 1 de l'item. Pour les modalités supérieures, on suppose que l'écart entre deux modalités adjacentes est le même pour tous les items. Il y a donc dans ce modèle des paramètres de seuils (« thresholds ») et qui sont supposés identiques d'un item à un autre.

Soit une échelle où chaque item possède m modalités de réponse. La probabilité p_{ijk} qu'a l'individu i de choisir la modalité de réponse k à l'item j connaissant son niveau de trait latent θ_i et le niveau de difficultés δ_j de l'item j et les seuils de difficultés $\tau_2, \dots, \tau_m$ des items est :

$$p_{ijk} = P\left(X_{ij} = k \middle| \theta_n,\ \delta_j, \tau_2, \dots, \tau_m\right) = \frac{\exp(k(\theta_i - \delta_j) - \sum_{l=2}^{k} \tau_l)}{1 + \sum_{h=1}^{m} \exp(h(\theta_i - \delta_j) - \sum_{l=2}^{m} \tau_l)}$$

4.2.2. Modèles de la famille de Lord

Les modèles de la famille de Lord sont des modèles « plus souples » que ceux de la famille de Rasch, avec un, deux ou trois paramètres par item. Ces modèles permettent généralement une meilleure adéquation aux données mais n'ont pas les mêmes propriétés que les modèles de la famille de Rasch.

Le premier paramètre de ces modèles reste celui de la difficulté de l'item. Le second paramètre correspond à un paramètre de discrimination ou aussi appelé pouvoir discriminant. Enfin, un paramètre de pseudo-chance peut être ajouté au modèle mais ce paramètre est davantage utilisé pour les tests de type QCM comme pour les tests de mathématiques que pour les études de QdV. En effet, dans le domaine de la QdV, on ne peut pas considérer qu'il y ait une « bonne » réponse. De plus, il est peu susceptible que les patients répondent au hasard.

Les modèles de la famille de Lord pour item dichotomiques sont souvent appelés par le nombre de paramètres caractérisant l'item, comme OPLM pour « One-Parameter Logistic Model », 2-PLM pour « Two-Parameter Logistic Model » ou encore 3-PLM pour « Three-Parameter Logistic Model » (De Ayala, 2009).

- **Modèle à un paramètre pour items dichotomiques**

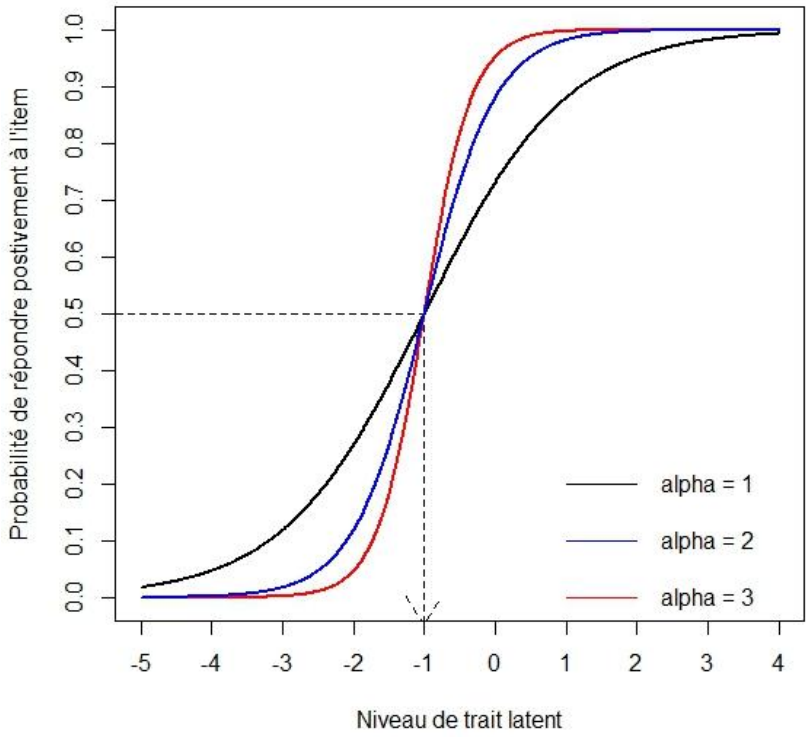
Le modèle « One- Parameter Logistic Model » (OPLM) est donc un modèle à un paramètre. Ce paramètre correspond au paramètre de difficulté de l'item. Ce modèle est adapté aux items dichotomiques. Selon ce modèle, chaque item est caractérisé par un paramètre de difficulté, mais aussi par une constante α correspondant au pouvoir discriminant de l'item. Ainsi, la probabilité p_{ij} qu'à l'individu i de choisir la réponse positive à l'item j, connaissant la valeur du trait latent, le pouvoir discriminant α des items ainsi que la difficulté δ_j de l'item j est donnée par l'équation suivante :

$$p_{ij} = P(X_{ij} = 1|\theta_i, \alpha, \delta_j) = \frac{\exp(\alpha(\theta_i - \delta_j))}{1 + \exp(\alpha(\theta_i - \delta_j))}$$

Le modèle OPLM correspond bien à un modèle à un paramètre (correspondant au paramètre de difficulté de l'item), puisque α est une constante et n'est pas susceptible de varier entre les items.

Le modèle OPLM et le modèle de Rasch sont sensiblement proches. Ces deux modèles requièrent que les items aient une valeur constante pour le pouvoir discriminant α. La différence entre les deux modèles réside au niveau des valeurs possibles de α. Le modèle de Rasch considère que le paramètre α est constant égal à un, pour tous les items, alors que dans le modèle OPLM, α peut ne pas être égal à un. Certains chercheurs peuvent considérer ces deux modèles comme étant identiques. Ces deux modèles sont de plus mathématiquement équivalents puisque les valeurs des paramètres obtenues pour l'un des modèles peuvent être transformées par un changement d'échelle appropriée en celles que l'on aurait obtenues par le second modèle (De Ayala, 2009). Cependant, les deux modèles (1-PLM et Rasch) représentent deux concepts philosophiques différents. En effet, le modèle 1-PLM, tout comme l'ensemble des modèles de la famille de Lord en général, se focalise sur l'ajustement des données au modèle. Ainsi, le modèle est construit de sorte à s'ajuster au mieux aux données, en ajoutant autant de paramètres que nécessaire. Inversement, les modèles de la famille de Rasch sont utilisés pour modéliser la variable d'intérêt, ce sont donc les données qui doivent s'adapter au modèle et non l'inverse (De Ayala, 2009).

A titre d'illustration, la Figure 5 représente la courbe caractéristique de trois items de même niveau de difficulté (δ = -1) et de pouvoir discriminant variable. Le paramètre de discrimination est lié à la pente de la courbe caractéristique de l'item. Il reflète ainsi la façon dont un item discrimine les individus le long du continuum latent. Plus le paramètre de discrimination est élevé, plus la pente de la courbe caractéristique de l'item est importante, et plus l'item contribuera au trait latent. Autrement dit, plus le paramètre de discrimination de l'item est élevé et plus l'item permettra de différencier les individus le long du continuum latent.



<u>Figure 5</u> : Courbe Caractéristique de trois items de même niveau de difficulté δ = -1 et de paramètre de discrimination α =1, 2 et 3 respectivement

- **Modèle à deux paramètres pour items dichotomiques**

Le modèle de Birnbaum ou aussi appelé 2-PLM (pour « 2 Parameter Logistic Model ») (De Ayala, 2009) est un modèle de la famille de Lord à 2 paramètres pour items dichotomiques. Les items sont caractérisés par un paramètre de difficulté de l'item mais aussi par le paramètre de discrimination. Contrairement au modèle OPLM, le paramètre de discrimination est susceptible de varier entre les items dans le modèle à deux paramètres.

Ainsi, la probabilité qu'a l'individu i de choisir la réponse positive à l'item j, connaissant la valeur du trait latent, le pouvoir discriminant α_j ainsi que la difficulté δ_j de l'item j est donnée par l'équation suivante :

$$p_{ij} = P(X_{ij} = 1|\theta_i, \alpha_j, \delta_j) = \frac{\exp(\alpha_j(\theta_i - \delta_j))}{1 + \exp(\alpha_j(\theta_i - \delta_j))}$$

Dans ce modèle, la statistique exhaustive du trait latent est la somme pondérée des réponses aux items. Le poids accordé à chaque item correspond au paramètre de discrimination de chaque item. Ceci permet de refléter l'importance relative de chaque item au regard du trait latent. De plus, différents profils de réponse correspondront à différentes valeurs de trait

latent. En effet, si deux individus obtiennent le même score observé, ils n'auront pas nécessairement la même valeur de trait latent. Par exemple, considérons deux individus obtiennant un score X = 2 en répondant positivement à deux items sur quatre tel que chaque item a un coefficient de discrimination variable. Si ces deux individus n'ont pas répondu positivement aux mêmes items, alors leur niveau de trait latent estimé par le modèle à deux paramètres sera différent, et ce quel que soit les coefficients de difficulté des items. Le modèle 2-PLM reste néanmoins un modèle pour item dichotomique.

- **Modèle à trois paramètres pour items dichotomiques**

Il existe également un modèle de Lord à trois paramètres (3-PLM pour « 3 Parameter Logistic Model) (De Ayala, 2009). Les deux premiers paramètres caractérisant l'item reste le paramètre de difficulté et le paramètre de discrimination. Le troisième paramètre correspond à un paramètre de pseudo-chance. Il caractérise le fait que les individus à faible capacité θ sont toujours susceptibles de donner une réponse positive (réponse « correcte ») à un item X donné en répondant au hasard. Pour un tel modèle, un paramètre de pseudo-chance non nul est reflété par une asymptote inférieure strictement positive, autrement dit :

$$\lim_{\theta \to -\infty} P(X = 1) > 0$$

Autrement dit, la probabilité qu'a un individu avec un niveau de trait latent infiniement bas de répondre positiement à l'item j est strictement positive, égale au paramètre de pseudo-chance χ_j de l'item j.

Ainsi, la probabilité p_j^* qu'à un individu i de répondre positivement à l'item j de paramètre de pseudo-chance χ_j, selon le modèle 3-PLM est :

$$p_j^* = p_j + \chi_j(1 - p_j)$$

où p_j est la probabilité obtenue par un modèle 2-PLM.

Si le niveau de trait latent de l'individu est très faible (tend vers -∞), alors la probabilité pour cet individu de répondre positivement à l'item selon le modèle 2-PLM tend vers 0 et donc, la probabilité pour cet individu de répondre positivement à l'item selon le modèle 3-PLM tend vers χ_j.

Cette équation peut se réécrire sous la forme suivante :

$$p_j^* = \chi_j + (1 - \chi_j)p_j$$

Enfin, la probabilité qu'a l'individu i de choisir la réponse positive à l'item j, connaissant la valeur du trait latent θ_i de l'individu, le pouvoir discriminant α_j ainsi que la difficulté δ_j de l'item j et le paramètre de peudo-chance χ_j de l'item j est donnée par l'équation suivante :

$$p_{ij} = P(X_{ij} = 1|\theta_i, \alpha_j, \delta_j, \chi_j) = \chi_j + \left(1 - \chi_j\right)\frac{\exp(\alpha_j(\theta_i - \delta_j))}{1 + \exp(\alpha_j(\theta_i - \delta_j))}$$

Cette probabilité est obtenue en remplaçant p_j par son expression, selon le modèle 2-PLM.

Comme nous l'avons déjà mentionné, ce modèle n'est pas adapté à l'analyse de la QdV. En effet, dans le domaine de la QdV, la notion de réponse « correcte » ou « incorrecte » n'est pas employée. De plus, il parait également peu envisageable qu'un individu réponde au hasard à un item.

- **Modèles à deux paramètres pour items polytomiques**

Il existe également deux modèles de la famille de Lord adaptés aux items polytomiques ordinaux : le modèle à crédit partiel généralisé (« Generalized Partial Credit Model », GPCM) et le modèle à réponse graduées (« Graded Response Model », GRM) (De Ayala, 2009; Samejima, 1969; Van der Linden & Hambleton, 1997). Ces deux modèles relâchent également l'hypothèse de paramètres de discrimination égaux entre tous les items.

Le modèle GPCM proposé par Muraki en 1992 est basé sur le modèle PCM dans lequel un paramètre de discrimination de l'item est ajouté (de la même façon que pour le modèle 2-PLM pour items dichotomique) (De Ayala, 2009). Ainsi, la probabilité de choisir la modalité de réponse k à l'item j à m_j modalités de réponse, connaissant le niveau de trait latent θ_i de l'individu i, le paramètre de discrimination α_j de l'item j et les niveaux de difficultés $\delta_{j,l}$ de l'item j, avec $l = 1, \dots, m_j$, et m_j modalités de réponse est :

$$P\left(X_{ij} = k \middle| \theta_i, \alpha_j\ \delta_{j,1}, \ldots, \delta_{j,m_j}\right) = \frac{exp\left(\sum_{l=1}^{k} \alpha_j(\theta_i - \delta_{j,l})\right)}{\sum_{c=1}^{m_j} exp\left(\sum_{l=1}^{c} \alpha_j(\theta_i - \delta_{j,l})\right)}$$

Ce modèle, comme le modèle PCM, est applicable pour des items avec un nombre variable de modalités de réponse.

Le modèle GPCM, tout comme le modèle PCM, est construit sur la base du modèle dichotomique (modèle de Rasch pour le PCM, modèle de Birnbaum (2-PLM) pour le GPCM). L'approche utilisée est celle d'une suite de modèles dichotomiques régissant la probabilité de choisir une modalité de réponse donnée plutôt que la modalité adjacente. L'approche utilisée par Samejima en 1969 pour développer le modèle GRM repose sur un concept différent (Samejima, 1969). Dans son approche, Samejima considère qu'il y a une limite au-dessus de laquelle on s'attend à ce qu'un individu donné obtienne un certain score donné plutôt qu'un score inférieur. Il s'agit donc ici d'un modèle cumulatif.

Le modèle GRM définit ainsi la probabilité qu'un individu i choisisse une certaine modalité de réponse k à un item j ou une modalité supérieure plutôt qu'il choisisse une modalité inférieure à la modalité k. Selon le modèle GRM, la probabilité d'obtenir un score x_j ou plus à l'item j est donc:

$$p_{x_j}(\theta) = \frac{\exp\left(\alpha_j\left(\theta\ - \delta_{x_j}\right)\right)}{1 + \exp\left(\alpha_j\left(\theta\ - \delta_{x_j}\right)\right)}$$

où θ est le trait latent, α_j le paramètre de discrimination de l'item j, δ_{xj} est le paramètre de localisation de la catégorie limite (catégorie seuil) pour le score x_j avec $x_j = [0,1, \ldots, m_j]$.

Il existe également des modèles IRT développés pour les items construits sur une échelle nominale. Ce type d'échelles étant rarement employé dans les instruments de mesures de QdV, ces modèles ne seront donc pas présentés ici. L'ensemble des modèles présentés ici sont résumés dans le Tableau 3, hormis le modèle 3-PLM n'étant pas utilisé dans le domaine de la QdV.

<u>Tableau 3</u>: Caractéristiques des principaux modèles IRT utilisés pour la validation des propriétés psychométriques des questionnaires

Modèle	Famille	Format de réponse à l'item	Paramètre de	
			Discrimination	Difficulté
Modèle de Rasch	Rasch	dichotomique	Egal à 1 pour tous les items	Varie entre les items
1-Parameter Logistic Model	Lord	dichotomique	Identique entre les items	Varie entre les items
2-Parameter Logistic Model	Lord	dichotomique	Varie entre les items	Varie entre les items
Partial Credit Model	Rasch	polytomique	Egal à 1 pour tous les items	Varie entre les items
Rating Scale Model	Rasch	polytomique	Egal à 1 pour tous les items	Distance entre les seuils constante entre les items
Graded Response Model	Lord	polytomique	Varie entre les items	Varie entre les items
Generalized Partial Credit Model	Lord	polytomique	Varie entre les items	Varie entre les items

Les modèles IRT présentés ici, en particulier les modèles de la famille de Rasch, permettent de valider les propriétés psychométriques des questionnaires en complément des analyses CTT. Ils permettent d'une part de vérifier que les items d'une même dimension ne sont pas redondants et d'autre part qu'ils sont bien adaptés à la population étudiée (Tennant & Conaghan, 2007). Ils permettent également de vérifier que les items sont bien construits, i.e. que les paramètres de difficulté des modalités de réponses des items sont bien ordonnés et que chaque modalité de réponse est bien représentée le long du trait latent, i.e. que chaque modalité de réponse est utile au regard de l'item. Une représentation graphique de cette propriété peut être obtenue à l'aide des ICC. Des recommandations sur la méthodologie d'analyse selon les modèles de la famille Rasch ont été proposées par Tennant *et al.* (Tennant & Conaghan, 2007). Les modèles de la famille de Lord sont globalement moins utilisés que ceux de la famille de Rasch pour étudier la validité des questionnaires, en particulier en raison des propriétés fondamentales des modèles de la famille de Rasch (Cappelleri *et al*, 2014).

Les modèles IRT sont également utilisés pour réduire des échelles de QdV. La sélection des items du questionnaire QLQ-C30 pour obtenir la version réduite QLQ-C15PAL a ainsi été réalisée selon des modèles IRT (Petersen *et al*, 2006).

De plus, les modèles IRT peuvent aussi être utilisés lors de la construction des items, pour sélectionner des items, revoir la formulation des items ainsi que pour le scoring. A titre d'illustration, le questionnaire de QdV BREAST-Q spécifique de la reconstruction mammaire a été développé selon le modèle RSM de la famille des modèles de Rasch (Pusic *et al*, 2009).

La sélection des items a été effectuée selon les résultats des analyses IRT. Le nombre de modalités de réponse par item a également été étudié selon les modèles IRT en étudiant les courbes caractéristiques des items : chaque modalité de réponse devait être importante au regard du trait latent, et les paramètres de seuil (« threshold ») ne devaient pas être inversés. L'adéquation au modèle de Rasch devait par ailleurs être respectée. Enfin, le scoring de ce questionnaire se fait également selon le modèle de Rasch, i.e. en estimant les paramètres de capacité des individus selon le modèle de Rasch. Le principal inconvénient de cette méthode de scoring est qu'elle nécessite un logiciel statistique pour l'estimation des scores.

Différentes études ont comparé la méthode classique de calcul des scores (par sommation des items) à la méthode basée sur les modèles IRT (McHorney *et al*, 1997; Norquist *et al*, 2004; Petersen *et al*, 2005). Il a ainsi été démontré que l'utilisation des modèles IRT pour le scoring pourrait permettre de gagner en sensibilité, particulièrement lorsque les items constituant l'échelle varient en terme de coefficients de difficulté (Norquist *et al*, 2004). Lorsque le questionnaire a été développé selon un modèle d'IRT, il parait naturel que le scoring soit effectué selon ce modèle plutôt que de procéder à un scoring selon la méthode classique (Pusic *et al*, 2009). En revanche, Petersen *et al.* n'ont rapporté aucun gain de la méthode IRT pour le scoring du questionnaire EORTC QLQ-C30 contrairement à la méthode classique (Petersen *et al*, 2005). Ceci peut s'appliquer par le faible nombre d'items par dimension et par la construction de l'échelle selon la méthode classique.

Enfin, les modèles IRT ont également été proposés pour déterminer la DMCI (Rouquette *et al*, 2014). L'objectif était d'investiguer la capacité des modèles IRT pour déterminer la DMCI tout en essayant de résoudre le problème de la dépendance de la DMCI au niveau du score à l'inclusion. Cependant, les auteurs n'ont pas montré d'apport substantiel des modèles IRT dans la détermination de la DCMI.

4.3. Fonctionnement différentiel de l'item

- **Définition**

La QdV dépend des attentes et espérances de santé du patient. Celles-ci peuvent varier entre différents groupes de patients. Si la perception d'un item diffère entre plusieurs groupes de patient, on parle de fonctionnement différentiel de l'item (« Differential Item Fonctioning »,

DIF) (De Ayala, 2009; Teresi & Fleishman, 2007; Zumbo, 1999). A titre d'illustration, supposons qu'un fonctionnement différentiel d'un item issu d'un questionnaire existe entre les hommes et les femmes. Cela signifie qu'à niveau égal de QdV sous-jacent, les hommes ne répondront pas de la même manière que les femmes à cet item. Une distinction est faite entre un DIF « uniforme » et « non uniforme » (Teresi & Fleishman, 2007). Un DIF uniforme indique que le fonctionnement différentiel est dans la même direction le long du trait latent, i.e. que pour chaque niveau de trait latent, la probabilité de choisir une modalité de réponse donnée pour un item est systématiquement plus faible (ou plus élevée) pour un groupe d'individus donné que pour un autre groupe. Un DIF non uniforme peut être vu comme une interaction significative entre la variable indicatrice du groupe et la capacité des individus. A titre d'illustration, considérons deux groupes d'individus A et B. Les individus à faible niveau de QdV dans le groupe A choisiront systématiquement des modalités de réponses plus basses que ceux du groupe B ayant le même niveau de QdV, alors que l'inverse est observé pour les individus à niveau de QdV élevé. Ces DIF « uniforme » et « non uniforme » correspondent à une différence au niveau des références internes des patients.

Même si la définition de la QdV est propre à chaque patient, cela ne doit pas affecter leurs réponses aux items, à niveau de QdV égal. Les items des questionnaires ne doivent donc pas présenter de DIF. Pour avoir une mesure objective, les paramètres du modèle doivent être indépendants des personnes qui y répondent : il s'agit de la propriété d'objectivité spécifique. A partir du moment où ce principe est violé, on parle de DIF. Si un tel DIF existe, il s'agit d'un biais de mesure et l'item affecté par le DIF ne peut pas en théorie être considéré comme valide. Ce DIF est constant au cours du temps. Si un DIF est démontré pour un item d'un questionnaire, il serait nécessaire de revoir la question posée et par exemple, de séparer cette question en deux pseudo-items selon l'appartenance de l'individu au groupe donné.

Les DIF les plus souvent rencontrés concernent l'âge, le sexe (Teresi *et al*, 2007), la région d'appartenance (Hardouin *et al*, 2012), ainsi que l'ethnie (Pagano & Gotay, 2005). L'adaptation d'un questionnaire dans une autre langue que celle d'origine requiert également une certaine prudence et nécessite une adaptation transculturelle plutôt qu'une simple traduction (Petersen *et al*, 2003). La présence éventuelle de ces DIF devrait donc être étudiée de façon systématique lors de l'adaptation d'un questionnaire, ce qui peut nécessiter d'avoir à la fois des données sur les patients de la langue d'origine et sur les patients de la langue dans laquelle le questionnaire est en cours d'adaptation.

- **Méthodes statistiques pour caractériser la présence de DIF**

Différentes méthodes statistiques ont été proposées pour investiguer la présence d'un fonctionnement différentiel de l'item (Hardouin *et al*, 2012; Scott *et al*, 2010; Scott *et al*, 2006). On distingue en particulier trois méthodes : le test du χ^2 de Mantel-Haenszel, les modèles de régression logistique et les modèles IRT.

Le test du **χ^2 de Mantel-Haenszel** est une méthode assez simple dont le test peut être effectué à la main. Cette méthode consiste à réaliser un tableau de contingence entre les sous-groupes de patients constitués et les réponses données à l'item considéré. On calcule alors les effectifs attendus de patients ayant choisi chaque modalité de réponse dans chaque groupe en l'absence de DIF. Un test du χ^2 de Mantel-Haenszel est alors appliqué pour examiner la différence entre les valeurs observées et celles attendues. Les Odds Ratio obtenus d'après ce test permettent d'estimer la magnitude du DIF : un Odds Ratio égal à 1 signifie l'absence de DIF, supérieur à 1 un DIF en faveur du 1er groupe et inférieur à 1 un DIF en faveur du 2nd groupe. Un DIF en faveur du groupe 1 signifie que les patients de ce groupe auront tendance à choisir des modalités de réponse plus élevées à l'item comparativement aux patients du 2nd groupe. Ce test a l'avantage d'être simple à réaliser. Cependant, seulement deux groupes de patients peuvent être considérés et seule la présence de DIF uniforme peut être détectée (Teresi, 2006).

Un **modèle de régression logistique** peut également être appliqué pour tester la présence de DIF (Scott *et al*, 2010; Scott *et al*, 2006). Ces modèles sont plus flexibles que le test du χ^2 de Mantel-Haenszel et sont également facile d'application. L'idée de base de cette approche est la même que pour le χ^2 de Mantel-Haenszel : examiner si pour chaque valeur observée du score global, l'item fonctionne de façon constante. Un modèle de régression logistique ordonnée est alors appliqué. La réponse à l'item est la variable explicative tandis que le score S obtenu pour l'échelle et la variable G indicatrice du groupe sont les variables dépendantes.

Le modèle s'écrit alors :

$$\log\left(\frac{P(X \leq i \,|S,G)}{1 - P(X \leq i \;\; S,G)}\right) = \; \beta_0 + \; \beta_1 \, S + \; \beta_2 \, G$$

où $P(X \leq i \,|S,G)$ est la probabilité de choisir une modalité de réponse inférieur ou égal à i connaissant la valeur du score S à l'échelle et le groupe d'appartenance G. Tout comme le test du χ^2 de Mantel-Haenszel, cette méthode ne peut détecter que la présence de DIF uniforme (Teresi, 2006).

Enfin, les **modèles IRT** permettent également de tester la présence de DIF (Hardouin *et al*, 2012; Langer *et al*, 2008; Pagano & Gotay, 2005). Un DIF représente une perception de l'échelle différente entre les sujets. Un DIF au niveau des références internes des individus peut ainsi être mis en évidence par un coefficient de difficulté de l'item qui diffère entre deux ou plusieurs groupes de patients. Un DIF uniforme correspond ainsi à un coefficient de la difficulté globale de l'item plus élevé ou plus faible pour un groupe d'individus comparativement à un autre groupe. Un DIF non uniforme correspond à un coefficient de difficulté pour une modalité de réponse donnée à un item plus élevé ou plus faible pour un groupe d'individus comparativement à un autre groupe. Cette méthode à l'avantage de pouvoir détecter la présence de DIF non uniforme. Cependant, elle requiert un bon ajustement au modèle d'IRT, un échantillon assez grand ainsi qu'un logiciel statistique adapté (Millsap, 2006).

5. Evaluation de la QdV dans les essais cliniques en cancérologie

5.1. Rationnel de la mesure de la QdV

Comme nous l'avons déjà mentionné dans le paragraphe 2, la QdV a été reconnue comme second critère de jugement principal par l'ASCO et la FDA en l'absence d'effet sur la survie globale (Beitz *et al*, 1996). La QdV est désormais intégrée dans la majorité des essais cliniques et particulièrement dans les essais de phase III (Burris *et al*, 2013; Kabbinavar *et al*, 2008; Stockler *et al*, 2014). Ainsi, en absence d'effet sur la survie globale ou sur tout autre critère de jugement principal, les résultats des données de QdV peuvent permettre de conclure à la supériorité ou non du nouveau traitement vis-à-vis du traitement de référence.

Dans cette optique, certains auteurs recommandent d'inclure la mesure de la QdV de façon systématique dans les essais cliniques de phase III en cancérologie (Osoba, 1992). Cependant, une majorité de chercheurs considèrent que la QdV ne doit être incorporée que dans certains types d'essais cliniques (Gotay *et al*, 1992; Moinpour *et al*, 1989).

Selon l'EORTC, l'évaluation de la QdV dans les essais cliniques en cancérologie est particulièrement importante lorsque (de Haes *et al*, 2000):

- Le pronostic des patients de l'étude n'est pas favorable ;
- L'impact attendu du traitement sur la QdV est important ;
- L'impact attendu du traitement sur la survie globale est petit.

Ainsi, il y a un large consensus indiquant que dans ces situations la QdV doit être considérée comme un critère de jugement majeur de l'essai clinique (Gotay *et al*, 1992).

Fayers *et al.* mettent également en évidence quatre situations où l'évaluation de la QdV semble être justifiée dans un essai clinique de phase III (Fayers & Machin, 2007):

- Les essais pour lesquels on s'attend à ce que le nouveau traitement ait un faible impact sur les critères cliniques tels que la survie à long terme ou la réponse tumorale et où un gain minime au niveau du critère de jugement principal doit être mis en parallèle avec les éventuels impacts négatifs sur la QdV du patient que pourraient engendrer un traitement plus intensif ;

- Les essais d'équivalence, où on s'attend à ce que l'évolution de la maladie soit similaire dans les deux bras de traitement mais où un bénéfice au niveau de la QdV est attendu. Dans ce type d'essais, la QdV doit être considérée comme un critère de jugement majeur ;
- Les essais pour lesquels il est prévu que le traitement améliore la QdV du patient. Ceci inclue en particulier les essais en situation palliative. Dans ces études, la QdV doit également être considérée comme un critère de jugement majeur ;
- Et les essais incluant une évaluation médico-économique contrebalançant le coût contre le gain potentiel de QdV.

De plus, bien que les recommandations pour l'évaluation de la QdV dans les essais cliniques concernent généralement les essais de phase III, la QdV est également régulièrement mesurée dans les essais de phase II (Garsa *et al*, 2013; Gontero *et al*, 2013; Mustea *et al*, 2013). La QdV peut également être intégrée à titre exploratoire dans les essais de phase I (Rouanne *et al*, 2013; Stephenson *et al*, 2013). Les essais de phase I ont pour objectif de déterminer la dose maximale tolérée et la dose recommandée pour les essais de phase II (Stephenson *et al*, 2013; Verweij *et al*, 2010). Cette dose est généralement déterminée à partir des évaluations de toxicités selon la grille « National Cancer Institute Common Terminology Criteria for Adverse Events » (NCI-CTC AE) (NCI, 2006). Or, l'avis et le ressenti du patient pourrait être informatif et en soit la mesure de la QdV ou de PROs centrés sur les symptômes et toxicités pourrait compléter les informations classiques de toxicités (Postel-Vinay *et al*, 2009). Certaines études de phase I ont déjà intégrées des mesures de QdV dans leur design mais les données restent encore sous exploitées et globalement les analyses sont uniquement descriptive (Stephenson *et al*, 2013). Une autre alternative est en cours d'investigation: il s'agit du développement d'une version patient de la grille NCI CTC-AE (« Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events ») (Bruner *et al*, 2011; Hay *et al*, 2014).

5.2. Planification de la mesure de la QdV

Le protocole de recherche de l'étude doit inclure un certain nombre d'information primordiale pour mener à bien l'étude. Des checklist ont été proposées par différents auteurs pour l'écriture des protocoles d'essais cliniques incluant la QdV (Fairclough, 2010; Fayers &

Machin, 2007). Ainsi, selon Fairclough (Fairclough, 2010), le protocole de recherche doit inclure les éléments suivants :

- Le rationnel pour l'étude de la QdV et les aspects spécifiques de la QdV à mesurer ;
- Des objectifs de recherches et critères de jugements explicites ;
- Des stratégies pour minimiser l'exclusion des sujets de l'essai ;
- Le rationnel pour les temps d'évaluation et les règles d'arrêt de l'étude ;
- Le rationnel pour la sélection de l'instrument de mesure ;
- Les détails concernant le mode d'administration du questionnaire de QdV pour minimiser les biais et les données manquantes ;
- Et un plan d'analyse statistique.

5.2.1. Rationnel et Objectifs

Le rationnel de la mesure de la QdV doit être clairement précisé dans le protocole de l'étude. Ceci permettra de justifier les ressources sollicitées pour investiguer la QdV et contribuera au succès de l'étude. Ce rationnel est généralement fourni pour tous les autres critères de jugements de l'étude. Les investigateurs ayant l'habitude de recueillir des données de laboratoires et des évaluations radiologiques, une motivation minimale est nécessaire au recueil de telles données. En revanche, davantage de motivation est généralement nécessaire au recueil de données de QdV nécessitant une auto-évaluation par le patient. Le rationnel de la mesure de la QdV permet ainsi de faciliter le recueil de telles données (Fairclough, 2010).

La justification de la mesure de la QdV dépend des objectifs de l'étude. Ces objectifs doivent être clairement définis. Un objectif tel que « comparer la qualité de vie entre les deux bras de traitements » n'est pas suffisant ni assez explicite. Ainsi, l'objectif doit répondre aux questions suivantes (Fairclough, 2010) :

- Quels sont les concepts spécifiques qui seront utilisés pour évaluer l'intervention ?
    - S'il s'agit de la QdV, est-ce la QdV globale ou une ou plusieurs dimensions spécifiques ?
    - S'il s'agit d'un symptôme, est-ce sa sévérité, le temps jusqu'à amélioration, l'impact sur les activités quotidiennes, etc.

- Quel est le l'objectif de la mesure de la QdV ?
    - Prétendre à un bénéfice au niveau de la QdV ?
    - Preuve de la supériorité de l'intervention ?
    - Analyse exploratoire de potentiels impacts négatifs ?
    - Evaluation pharmaco-économique ?
    - Autre ?
- Quelle est la population ciblée?
- Quelle est la fenêtre temporelle d'intérêt ?
- Est-ce une étude à visée exploratoire ou confirmatoire ?

5.2.2. Sélection des patients

Les patients sélectionnés pour l'évaluation de la QdV doivent être, dans la mesure du possible, l'ensemble des patients sélectionnés pour les autres critères de jugements de l'étude (Fairclough, 2010). Ces recommandations ont été données par Gotay *et al.* pour des raisons à la fois pratiques et scientifiques (Gotay *et al*, 1992). Lorsque les critères de jugements sont évalués pour tous les patients inclus, la crédibilité et l'interprétation des résultats sont accrus (Fairclough, 2010; Gotay *et al*, 1992).

Néanmoins, le remplissage de questionnaires de QdV peut parfois ne pas être réalisable pour tous les patients. En effet, les patients doivent présenter des capacités physiques, cognitives et linguistiques suffisantes pour une auto-évaluation de la QdV. Dans certains cas, ces critères font partie des critères d'inclusion de l'étude. Dans le cas contraire, selon la population ciblée, certains patients pourraient être exclus de l'étude de la QdV. Si la proportion de patients exclus est négligeable, il n'y aura pas de conséquence au niveau de l'analyse. En revanche, si la proportion de patients exclus est élevée, cela risque d'engendrer un biais de sélection de la population pour la QdV et ainsi les résultats de l'étude de la QdV ne correspondront pas à ceux que l'on aurait obtenus sur la population initialement ciblée. Dans de telles situations, il serait préférable d'autoriser une hétéro-évaluation de la QdV en permettant que le questionnaire soit énoncé au patient par une tierce personne. Par exemple, en situation palliative, un hétéro-questionnaire tel que le QUAL-E pourrait être plus adapté qu'un auto-questionnaire puisque la majorité des patients présenteront des capacités physiques trop limitées pour pouvoir remplir un questionnaire.

5.2.3. ***Temps d'évaluation de la QdV***

Dans les essais cliniques en général, au moins trois évaluations de la QdV sont réalisées et requises : à l'inclusion dans l'étude (avant le début de l'administration du traitement), pendant le traitement et à la fin de l'étude après arrêt du traitement. Le nombre et le moment de passation des questionnaires dépendent à la fois des objectifs de l'étude et de la faisabilité. Néanmoins, un minimum de deux mesures est nécessaire : à l'inclusion et à la fin de l'étude ou après un arrêt prématuré du traitement (Wiklund, 2004).

Il est essentiel que la première évaluation de la QdV ait lieu non seulement avant le début du traitement, mais également avant la randomisation du traitement. En effet, la QdV étant un critère de jugement subjectif, la connaissance de l'attribution du traitement par le patient pourrait influencer son évaluation de la QdV (Brooks *et al*, 1998).

Durant le traitement, deux types de schémas d'évaluations peuvent être réalisés. Une évaluation « Event-driven design » sera réalisée si les schémas d'administration de traitements diffèrent entre les deux bras considérés. L'évaluation est généralement planifiée juste avant une intervention telle que la chirurgie ou le début de la radiothérapie. L'objectif de de ces évaluations est d'étudier l'effet du traitement sur la QdV à court terme. En revanche, lorsque l'objectif est d'étudier la QdV à long terme et que les deux types de prise en charge respectent le même schéma, une évaluation de type « Time-driven design » est réalisée. Dans ce type de design, les temps d'évaluation sont basés sur des intervalles de temps régulièrement espacés. Par exemple, tous les mois pendant trois mois, puis tous les trois mois pendant un an. Certaines études peuvent également être basées sur un mélange de ces deux types de schémas d'évaluation.

L'intervalle de temps entre les évaluations de la QdV, ainsi que le schéma et le nombre d'évaluations peuvent influencer le choix de l'approche statistique pour l'analyse des données.

La fréquence des évaluations dépend de la localisation cancéreuse, de la situation thérapeutique ainsi que de la probabilité d'observer un changement de la QdV durant cette période. Les évaluations doivent être assez nombreuses pour pouvoir capturer un changement significatif d'un point de vue clinique (soit la MCID). En revanche, des évaluations en trop grand nombre ou trop rapprochées dans le temps risqueraient également d'ennuyer ou de lasser le patient. Si un changement de QdV est attendu dans les premiers mois de l'étude par

exemple, alors des évaluations plus fréquentes de la QdV doivent avoir lieu durant cette période.

L'évaluation de la QdV est généralement planifiée jusqu'à la progression de la maladie et/ou l'arrêt du traitement. Pour des aspects pratiques et scientifiques, des recommandations doivent être développées concernant le suivi des patients qui n'ont pas suivi le protocole du traitement. En effet, un bras de traitement avec un fort taux de sorties d'étude prématurées pourrait paraître bénéfique puisque seuls les patients en meilleur santé seraient maintenus dans l'étude (Fairclough, 2010).

L'objectif de ces évaluations de la QdV au cours du temps est d'étudier l'impact du traitement et de la progression de la maladie sur la QdV du patient. Une étude longitudinale de ces données de QdV est donc requise.

5.2.4. Sélection du questionnaire de QdV

Le choix du questionnaire peut être crucial pour mener à bien une étude. Ce choix dépend des objectifs de l'étude et des caractéristiques de la population ciblée.

Une checklist pour le choix du questionnaire a été proposée par différents auteurs. A titre d'exemple, la checklist proposée par Fairclough *et al.* pour la sélection des instruments de mesure dans un essai clinique est la suivante (Fairclough, 2010) :

- Identification du concept à mesurer
- Est-ce que l'instrument mesure ce qu'il est censé mesurer ?
- Est-ce que l'instrument est pertinent pour l'objectif de l'étude ?
    - Dans quelle mesure l'instrument couvre bien tous les domaines importants qu'il est censé mesurer ?
    - Est-ce qu'un questionnaire générique ou spécifique est plus approprié ?
- Est-ce que l'instrument permettra de discriminer les sujets de l'étude et de détecter un changement ?
- Est-ce que les items sont appropriés aux sujets à tout moment de l'étude ?
- Est-ce que le format et le mode d'administration sont appropriés aux sujets et à l'essai ?
- Est-ce que l'instrument a été validé préalablement auprès de la population cible ?

- Si de nouveaux instruments ou items sont utilisés, quel est le rationnel pour leur utilisation ?

Le choix entre un questionnaire générique et un questionnaire spécifique dépend des objectifs de l'étude. Les questionnaires génériques ont été conçus pour être utilisés dans la population générale ou pour comparer des patients présentant des maladies très diverses. Dans un essai clinique en cancérologie, un questionnaire générique peut être utile si l'objectif est de comparer les données de QdV des patients de l'essai à celles d'autres patients issu de la population générale ou ayant une autre pathologie ou une autre localisation cancéreuse. Les questionnaires génériques sont le plus souvent utilisés dans le cadre d'études médico-économique telles que les QALYs (voir paragraphe 6.2.6).

Les questionnaires spécifiques du cancer ont été développés pour mesurer préférentiellement les domaines de QdV potentiellement impactés par la maladie ainsi que les symptômes et effets secondaires propres à cette maladie et à ses traitements. Ces questionnaires sont donc plus sensibles à un changement de QdV au cours du temps pour un patient donné ou pour détecter une différence de niveau de QdV entre deux groupes de patients donné (bras contrôle vs. bras expérimental par exemple) (Teresi, 2006). Si l'objectif de la mesure de la QdV dans l'essai est de détecter une telle différence, l'utilisation d'un questionnaire spécifique du cancer (et idéalement de la localisation cancéreuse) est préférable.

Le choix entre un questionnaire de l'EORTC et un questionnaire du groupe FACT se fait généralement selon les préférences et affinité des équipes associées à l'essai clinique. Différentes études ont comparé les questionnaires de QdV de l'EORTC à ceux du groupe FACT, en particulier pour les questionnaires généraux (Kemmler *et al*, 1999), mais également pour ceux spécifiques des cancers en situations avancées (Lien *et al*, 2011), ou des métastases osseuses par exemple (Popovic *et al*, 2012). Il s'avère que même si leur structure est totalement différente, il n'y a globalement pas de questionnaire préférable parmi ces deux groupes d'un point de vue de la validité et de la performance. Les études nord-américaines seront naturellement davantage portées vers les questionnaires FACT et les études européennes vers les questionnaires de l'EORTC.

De plus, si l'objectif est de cibler spécifiquement les symptômes propres à une localisation cancéreuse, un questionnaire spécifiquement développé pour cette localisation peut être plus adapté, comme le questionnaire EPIC pour le cancer de la prostate par exemple.

Cependant, ces nombreuses possibilités de questionnaires posent des problèmes de comparabilités des résultats entre les essais. Par exemple, il a été démontré que des résultats obtenus avec les deux questionnaires généraux EORTC QLQ-C30 et FACT-G ne peuvent être directement comparés (Kemmler *et al*, 1999). Il parait donc indispensable de proposer des recommandations pour le choix du questionnaire selon les essais cliniques afin de pouvoir comparer les résultats issus de plusieurs essais cliniques pour une même localisation cancéreuse, une même situation thérapeutique et des modalités de traitement similaires.

Plusieurs questionnaires de QdV ou de PROs peuvent également être retenus afin de capter au mieux les différents aspects de la QdV des patients potentiellement impactés par la maladie. Par exemple, il peut paraître pertinent de considérer un questionnaire de QdV spécifique du cancer comme le QLQ-C30 ainsi qu'un questionnaire de sexualité, ou d'anxiété et de dépression selon l'étude.

Néanmoins, il est important de ne pas surcharger le patient en lui proposant des questionnaires trop longs ou trop nombreux, en particulier dans les études où les patients sont en situation avancée de cancer voir en situation palliative. En effet, si le temps de remplissage des questionnaires est trop long, le patient risque de ne pas remplir intégralement les questionnaires ce qui entrainera l'occurrence de données manquantes.

Dans la mesure du possible, il est fortement recommandé de choisir un questionnaire dont la validation dans la langue de la population étudiée a déjà été effectuée dans une étude antérieure (Streiner & Norman, 2008). Dans le cas contraire, une validation des propriétés psychométriques du questionnaire doit être prévue en parallèle de l'étude principale.

Une checklist a été proposé par Fayers *et al.* (Fayers & Machin, 2007) pour le choix d'un questionnaire validé et adapté à la population d'étude adressant les différents points suivants :

- Documentation

1. Existe-t-il une documentation écrite à propos du questionnaire?
2. Existe-t-il un manuel de l'utilisateur?

- Développement

1. Est-ce que les objectifs et l'usage prévu de l'instrument sont clairement définis?
2. Existe-t-il une base conceptuelle claire pour les dimensions évaluées?
3. Est-ce que l'instrument a été développé selon des procédures rigoureuses? Est-ce que les résultats ont été publiés en détail? Ceci doit inclure toutes les étapes depuis l'identification des domaines et la sélection des items jusqu'à un test à large échelle.

- Validation

1. Est-ce que le processus de validation est compréhensible, et est-ce que les études de validations ont été menées avec une taille d'échantillon adéquate?
2. Est-ce que les dimensions validées correspondent aux concepts pertinents pour votre étude?
3. Existe-t-il une documentation attestant d'une validation adéquate?
4. Existe-t-il des preuves d'une reproductibilité et fiabilité adéquate des résultats?
5. Existe-t-il une preuve de sensibilité et de « responsiveness »? En quoi ces valeurs affectent la taille d'échantillon nécessaire pour votre étude?

- Population cible

1. Est-ce que l'instrument est approprié pour votre population cible? A-t-il été testé auprès de sujets de cette population (i.e. patients avec la même maladie, même stade, recevant un traitement similaire)?
2. Si votre population diffère de celle ciblée, est-ce raisonnable de s'attendre à ce que l'instrument soit applicable? Est-ce qu'un test supplémentaire est nécessaire pour le confirmer?
3. Est-ce que votre étude va inclure des sujets particuliers, tel que de jeunes enfants ou des adultes avec des handicaps cognitifs, pour lesquels le questionnaire pourrait être moins approprié?

- Faisabilité

1. Est-ce que la méthode d'administration est faisable?
2. Est-ce que les questions sont compréhensibles à la lecture ou est-ce qu'une aide est nécessaire?
3. Combien de temps est nécessaire pour compléter le questionnaire?
4. Est-ce que le questionnaire contient des items difficiles ou embarrassants?
5. Est-ce que le traitement du questionnaire est facile ou est-ce que des items requièrent un codage, tels que des échelles visuelles analogiques?
6. Si de multiples questionnaires doivent être utilisés, sont-ils compatibles entre eux?

- Langages et cultures

1. Est-ce que l'instrument a été testé et validé pour utilisation auprès de patients de niveau d'éducation, culturel et ethnique pertinent?
2. Est-ce qu'il existe des traductions valides qui couvrent vos besoins, présents et futurs?

3. Si des versions dans d'autres langues sont nécessaires, elles doivent être développées selon des procédures formelles de traduction forward et backward et testées auprès d'un certain nombre de patients qui ont aussi complété un questionnaire de debriefing.

- Scoring

1. Est-ce que la procédure de scoring est définie? Y a-t-il un score global pour la QdV globale?
2. Y a-t-il des questions globales pour évaluer la QdV globale?

- Interprétation

1. Existe-t-il des recommandations pour interpréter les échelles de scores?
2. Existe-t-il des données de référence ou d'autres recommandations pour l'estimation d'une taille d'échantillon lorsque l'on désigne un essai?
3. Y a-t-il une question globale ou une mesure globale de la QdV?
4. Existe-t-il ou est-il nécessaire de fournir une question ouverte à la fin à propos d' « autres facteurs affectant votre QdV qui n'a pas été mentionné plus haut »?
5. Est-ce que les effets secondaires des traitements sont couverts de façon adéquate?

5.2.5. Dimensions ciblées

Une fois que le choix du (ou des) questionnaire(s) a été fixé, il est nécessaire de cibler les dimensions d'intérêt premier (entre deux et cinq dimensions) afin de prévenir l'inflation du risque de première espèce. Ces dimensions doivent être clairement identifiées dans le protocole de l'étude. Les autres dimensions seront étudiées à titre exploratoire ou comme objectif secondaire.

Les dimensions prioritaires doivent être, dans la mesure du possible, des dimensions multi-items. En effet, l'évaluation des domaines de QdV d'intérêt premier sera d'autant plus précise que le nombre d'items est important.

Afin de prévenir l'inflation du risque de première espèce dû à la multiplicité des analyses, il est recommandé de faire un ajustement du risque de première espèce, en utilisant par exemple un ajustement de Bonferroni, consistant à diviser le risque de première espèce par le nombre de dimensions de QdV retenues. Cet ajustement peut être évité si la QdV est un critère secondaire de l'étude.

Lorsque plusieurs dimensions de QdV sont étudiées, les règles décisionnelles des critères de jugements multiples s'appliquent. Selon les situations, on pourra considérer que la mise en évidence d'au moins un résultat significatif est suffisante (« single sufficient rule »). Inversement, on pourra considérer que l'étude est significative si toutes les dimensions sont significatives.

Une procédure des tests fermés (« close testing ») peut également être réalisée (Marcus *et al*, 1976). Dans e cas, un test global est tout d'abord réalisé sur l'ensemble des n dimensions de QdV au niveau de signification α sans ajustement. Si le test s'avère significatif, tous les tests globaux intégrant n-1 dimensions de QdV seront réalisés. A l'étape suivante, les tests comprenant n-2 dimensions de QdV sont réalisés seulement si tous les tests incluant n-1 critères de jugement son significatif, et ainsi de suite jusqu'à a réalisation des tests n'intégrant qu'une seule dimension de QdV.

5.2.6. Nombre de sujets nécessaires

La détermination du nombre de sujets nécessaires est une étape primordiale de la conceptualisation de l'étude. La taille de l'échantillon est déterminée selon le critère de jugement principal. Si la QdV est l'objectif principal de l'étude, la taille de l'échantillon doit alors être calculée par rapport aux dimensions de QdV ciblées, en tenant compte de la DMCI, du taux d'erreur de type I et de la puissance souhaitée. Si plusieurs dimensions de QdV sont ciblées, alors un ajustement du taux d'erreur de type I doit être réalisé afin de tenir compte de la multiplicité des tests réalisés. Enfin, la détermination de la taille de l'échantillon dépendra de la méthode statistique considérée. A ce jour, aucun essai clinique ne considère la QdV comme critère de jugement principal. Cependant, elle pourrait être considérée comme co-critère de jugement principal ou critère de jugement composite avec un critère tumoral tel que la survie sans progression.

Pour les analyses longitudinales, un pourcentage de 5 ou 10% de perdu de vue ou de données manquantes est généralement pris en compte pour le calcul du nombre final de patients à inclure dans l'essai.

Si la QdV est un critère de jugement secondaire, alors la taille d'échantillon déterminée selon le critère de jugement principal peut ne pas être adaptée à une analyse longitudinale des données de QdV.

5.3. Déroulement de l'étude

5.3.1. Ordre des questionnaires et place des évaluations

L'ordre des questionnaires peut influencer les réponses. Cet aspect peut parfois être négligé lors de la mise en place de l'étude. Par exemple, si des questions sur les effets secondaires sont placées avant une question sur la QdV globale, alors il existera une forte corrélation entre la réponse à la dernière question et les effets secondaires. Si plusieurs questionnaires sont administrés, celui représentant le critère de jugement principal pour la QdV doit être présenté en premier. En effet, si le patient doit remplir plusieurs questionnaires avec un temps total de remplissage assez long, le patient risque de moins bien remplir les derniers questionnaires comparativement au premier. Pour ce qui est des questionnaires EORTC, le questionnaire QLQ-C30 doit être rempli avant ses modules spécifiques additionnels.

Le lieu et la fenêtre temporelle pour le remplissage peuvent également influencer le choix des réponses du patient. En effet, il a été démontré que si un patient complète le questionnaire à domicile il ne donnera pas nécessairement les mêmes réponses que s'il était à l'hôpital (Millsap, 2006). De plus, le remplissage du questionnaire pourra être différent selon que le patient le remplisse avant ou après un entretien avec son médecin (Noll & Fairclough, 2004).

5.3.2. Modalités de remplissage des questionnaires

Différents mode d'administration des questionnaires existent, comme le mode « papier-stylo », par téléphone ou lors d'un entretien avec un attaché de recherche clinique. Le gold standard pour l'évaluation et le mode de remplissage des questionnaires reste à ce jour le mode « papier-stylo ». Le mode d'administration des questionnaires peut influencer les résultats et le mode « papier-stylo » semble ainsi le moins susceptible de présenter des biais de mesure comme le biais de désirabilité sociale (Buskirk & Stein, 2008; Cheung *et al*, 2006; Weinberger *et al*, 1996).

Des outils informatisés sont également de plus en plus développés. Le groupe EORTC a par exemple développé un programme informatisé pour l'évaluation et l'analyse de la QdV (Holzner *et al*, 2012). Cette méthode permettrait d'avoir des questionnaires remplis correctement et sans données manquantes et serait en particulier utile dans le cadre d'une évaluation de la QdV en routine.

Enfin, si une aide au remplissage des questionnaires est autorisée dans le protocole de l'étude, le recours à cette aide doit pouvoir être renseignée dans le cahier d'observation de l'étude. L'identité de la personne ayant aidé au remplissage (médecin, attaché de recherche clinique, famille ou autre) doit également être renseignée.

5.3.3. Prévention des données manquantes

La prévention des données manquantes est importante dans tout essai clinique (Little *et al*, 2012). Pour ce qui est des questionnaires de QdV, les données manquantes au sein d'un questionnaire ou les questionnaires manquants peuvent entraver l'analyse des données. Ainsi, il est essentiel d'expliquer au patient l'importance de chaque réponse. Le rôle de l'attaché de recherche clinique est aussi essentiel dans cette prévention des données manquantes. Il est également important de ne pas surcharger le patient avec des questionnaires trop longs, trop nombreux ou un intervalle de temps entre les évaluations trop court.

Malgré ces précautions, il est possible que le patient ne puisse pas remplir le questionnaire à un temps de mesure donné. Il est alors important de recueillir les motifs de non remplissage des questionnaires : le patient était-il trop malade ou trop fatigué pour remplir le questionnaire ? Le médecin pensait-il que le patient était trop malade pour remplir le questionnaire ? Le patient a-t-il déménagé ? Etc.

5.4. Recommandations pour le report des résultats

Des recommandations ont été proposées pour la mesure de la QdV (Osoba, 2011) et le report des résultats de QdV (Calvert *et al*, 2013; Calvert *et al*, 2011) dans les essais cliniques en cancérologie. Les informations à reporter dans toute étude analysant la QdV selon Fayers *et al.* (Fayers & Machin, 2007) sont résumées dans le Tableau 4 suivant.

<u>Tableau 4</u> : Informations à reporter dans les études intégrant la QdV selon Fayers et al.

Section/thème	Description
Introduction	
	Décrire les objectifs de l'étude
	Décrire brièvement l'épidémiologie et l'éthologie de la maladie et le rationnel de la mesure de la QdV dans cette étude
	Les hypothèses de l'étude concernant l'évaluation de la QdV doivent être indiquées
	La définition de la QdV doit être décrite
Méthodes	
Participants	Les critères d'éligibilité des participants et le cadre dans lequel les données ont été collectées
Sélection de l'instrument de QdV	Rationnel pour le choix du questionnaire de QdV, incluant les détails de la validation des propriétés psychométriques
Nombre de sujets nécessaires	Calcul de la taille de l'échantillon, et la définition de la DMCI
Résultats	Indiquer les dimensions de QdV prioritaires
Temps et mode d'administration des questionnaires	Reporter les temps d'évaluation planifiés
	Décrire la méthode d'administration
Données	Préciser les méthodes pour l'imputation des items manquantes et/ou questionnaires manquants
Méthodes statistiques	Définition du seuil de signification statistique
	Description des méthodes statistiques utilisées pour comparer les groupes concernant le critère de jugement principal et les analyses supplémentaires, telles que les analyses en sous-groupes et analyses ajustées
	Tout ajustement pour les comparaisons multiples doit être indiqué
Résultats	
Flux des participants	Flux des patients à chaque étape (un diagramme est fortement recommandé).
	Pour chaque groupe, reporter le nombre de patients randomisés, recevant le traitement, ayant complété le protocole de l'étude, et analysé pour le critère de jugement principal. Décrire les déviations au protocole de l'étude
Recrutement	Dates définissant les périodes de recrutement et de suivi
Données à l'inclusion	Les données cliniques, démographiques et de QdV à l'inclusion par bras de traitement doivent être reportées sous forme de tableau
Nombre de patients analysés	Pour chaque groupe, donner le nombre de patients inclus dans chaque analyse et si l'analyse était en intention de traiter
Résultats et estimation	Donner un résumé des résultats par groupe et pour chaque dimension, avec l'Effect Size estimé et sa précision (intervalle de confiance à 95%)
Analyses ancillaires	Résultats de toute autre analyse réalisée, incluant les analyses en sous-groupes et analyses ajustées, indiquant les analyses pré spécifiées et les analyses exploratoires
Discussion	
Interprétation	Interprétation des résultats de QdV, tenant compte des hypothèses de l'étude, source de biais potentiel ou d'imprécision et les dangers associés avec les analyses multiples
	Un résumé des résultats cliniques doit être reporté en parallèle des résultats de QdV permettant une interprétation équilibrée des résultats
Généralisation	Généralisation (validité externe) des recherches de l'étude

Pour les essais cliniques, des recommandations ont été proposées sur les éléments à rapporter dans les publications sous la forme de critères CONSORT (« Consolidated Standards of Reporting Trials ») (Schulz *et al*, 2010). Ces recommandations sont suivies à travers le monde et sont demandées dans la plupart des revues lorsqu'il s'agit de publier les résultats d'un essai clinique. Une extension de ces critères CONSORT à la QdV a été récemment proposée par Calvert *et al*. (Calvert *et al*, 2013) : il s'agit des critères CONSORT PRO. Contenant des éléments de la Checklist de Fayers *et al*., il est fortement conseillé de suivre ces recommandations pour les analyses de QdV dans les essais cliniques afin d'obtenir une publication de qualité.

6. Analyse longitudinale des données de QdV

L'analyse longitudinale des données de QdV pose certaines difficultés méthodologiques dues au caractère subjectif et dynamique de la QdV, à l'occurrence de données manquantes, à l'interprétation des résultats au regard du sens clinique, et à la nécessité de proposer une méthode statistique adaptée à la nature des données.

6.1. Définition de la population d'analyse

Avant de réaliser l'analyse longitudinale des données de QdV, il est nécessaire de définir la population d'analyse. Afin de pouvoir interpréter au mieux les résultats et d'éviter tout biais de sélection de patients, l'idéal serait que la population d'analyse de la QdV soit la même que celle du critère de jugement principal (Fairclough, 2010). Généralement, tous les patients randomisés, quel que soit le respect des critères d'éligibilité sont analysés : il s'agit de l'analyse en intention de traiter. Dans le cadre de l'étude de la QdV, la présence d'au moins une mesure de QdV par patient est généralement nécessaire pour pouvoir réaliser l'analyse : il s'agit alors d'une analyse en intention de traiter modifiée. Ainsi, l'analyse peut être réalisée en sélectionnant les patients ayant au moins la mesure initiale de disponible. Dans tous les cas, cette population doit être clairement définie en amont de l'analyse.

6.2. Méthodes d'analyse longitudinale

L'analyse de mesures répétées de QdV pourrait être réalisée en utilisant une série de test pour données indépendantes (par exemple, à l'aide d'un test T de Student, une ANOVA ou une MANOVA) pour tester la différence de QdV entre deux bras de traitement à chaque temps de mesure. Il s'agirait donc d'une analyse temps par temps. Cette méthode est simple mais a de nombreuses limites (Everitt, 2001) :

- Les données de QdV entre deux mesures ne sont pas indépendantes, donc l'interprétation des résultats peut-être difficile ;
- La multiplication des tests effectués à chaque temps de mesure peut entrainer une inflation du risque alpha, donc des résultats significatifs pourront être uniquement dus au hasard ;
- La perte de l'information relative aux changements de QdV intra-sujets au cours du temps.

Une analyse de l'évolution de la QdV entre deux temps de mesure pour deux bras de traitement pourrait être réalisée selon une méthode d'analyse transversale de type analyse de variance (ANOVA), soit en considérant uniquement les mesures au temps T2 (et en considérant que les niveaux de QdV à l'inclusion sont identiques selon les deux bras), soit en ajustant sur les mesures à T1 (analyse de covariance). Cependant, cette analyse ne permet pas d'explorer l'évolution de la QdV avec plus de deux mesures. Or, aux moins trois mesures sont généralement planifiées dans les essais cliniques en cancérologie.

Une analyse longitudinale de ces données est donc préférable puisque elle permettra de tenir compte de la corrélation entre les mesures d'un même patient au cours du temps. Différentes méthodes d'analyses statistiques ont été proposées pour analyser les données de QdV longitudinales (Bacci, 2008; Bonnetain *et al*, 2010; Cnaan *et al*, 1997; Douglas, 1999; Fairclough, 2010; Hamidou *et al*, 2011). Si l'analyse longitudinale de la QdV est l'objectif de jugement principal de l'étude, la taille de l'échantillon doit être calculée en tenant compte de la méthode statistique retenue ce qui peut être complexe selon les méthodes.

6.2.1. Aire sous la courbe

L'aire sous la courbe (« Area Under the Curve », AUC) est une technique assez ancienne qui permet de résumer simplement des mesures de QdV longitudinales (Fayers & Machin, 2007). Cette méthode peut s'appliquer sur données complètes ou incomplètes.

Lorsque l'on représente graphiquement les données de QdV individuelles d'un patient (Figure 6), l'AUC correspond à l'aire entre les mesures de QdV d'un patient et l'axe des abscisses représentant un niveau de QdV (ou de symptôme) le plus faible possible (soit égal à 0).

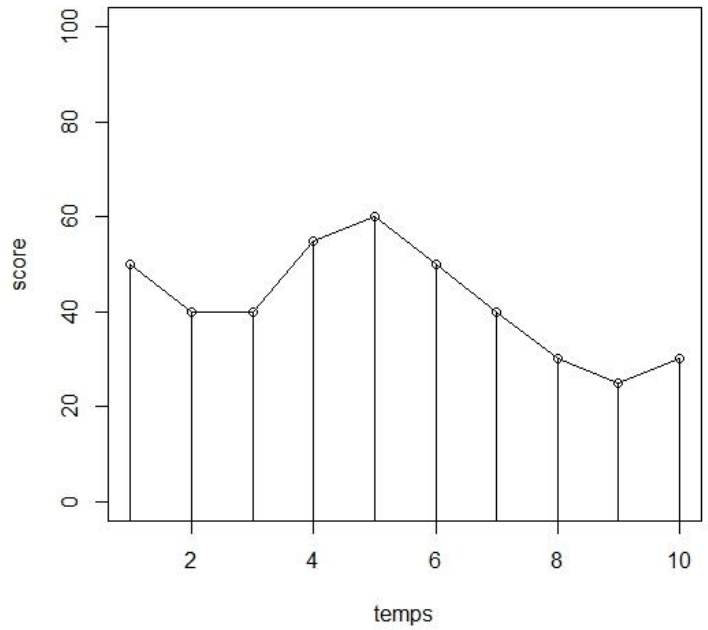


Figure 6 : Données de QdV individuelles pour un patient au cours du temps et estimation de son AUC selon la méthode du trapèze

La technique du trapèze est généralement utilisée pour calculer l'AUC. Elle consiste à relier les mesures de QdV consécutives d'un même patient par une ligne droite, donnant ainsi l'allure de la courbe pour un patient. L'aire est ensuite calculée en sommant les aires sous la courbe entre chaque paire d'observation consécutives.

Par exemple, l'aire A_i entre deux évaluations de QdV consécutives à t_i et t_{i+1} pour un patient donné est :

$$A_i = \frac{(QdV_i + QdV_{i+1})(t_{i+1} - t_i)}{2}$$

où QdV_i et QdV_{i+1} représentent les scores de QdV du patient respectivement à t_i et t_{i+1}.

L'AUC d'un patient avec $n + 1$ évaluations de QdV est donc égal à :

$$AUC = \frac{(A_1 + .. + A_i + .. + A_n)}{t_{n+1} - t_1}$$

Le dénominateur correspond à la durée totale de suivi du patient, i.e. l'intervalle de temps entre la première et la dernière évaluation de QdV. Cette division finale permet de tenir compte des différences individuelles des périodes de suivi.

Le calcul de l'AUC se fait donc de façon individuel. On résume ainsi les données longitudinales par individu par une quantité unique. L'AUC peut être calculée même en présence de données manquantes. L'avantage de cette approche est que l'AUC obtenue par patient est davantage susceptible de suivre une distribution normale que les données de QdV originelles (Fayers & Machin, 2007). La moyenne des aires sous la courbe peut donc être calculée et comparée par bras de traitements en utilisant un test de Z, test de T, une ANOVA ou un modèle de régression. Un modèle d'analyse transversale est ainsi appliqué et non un modèle d'analyse longitudinale.

Cette méthode est donc assez simple. Elle permet de résumer des mesures répétées pour un même patient par une quantité unique. Cependant, il s'agit d'une méthode ancienne, souvent utilisée par le passé (Lydick *et al*, 1995), mais qui ne permet pas d'analyser en profondeur les données longitudinales des patients telle que la magnitude de l'effet traitement. Une méthode d'analyse longitudinale semble donc préférable.

6.2.2. Modèle linéaire à effets mixtes basé sur le score

Le modèle linéaire à effets mixtes basé sur le score est la méthode statistique la plus utilisée à ce jour pour l'analyse longitudinale de la QdV (Ferrandina *et al*, 2014; Mantegna *et al*, 2013; Pan *et al*, 2012; Schaake *et al*, 2013). Elle consiste à appliquer un modèle linéaire à effets mixte sur le score observé et calculé à chaque temps de mesure (Cnaan *et al*, 1997). Des effets fixes et des effets aléatoires peuvent être introduits dans le modèle.

Le modèle de base considère un effet fixe temps et un effet fixe bras de traitement. L'effet bras correspond à l'écart entre les deux bras de traitement au temps initial et est considéré constant au cours du temps. Dans un essai clinique randomisé, les deux bras de traitement

sont supposés présenter le même niveau de QdV initial. Un effet d'interaction peut également être introduit dans le modèle afin de tester si le niveau de QdV évolue différemment entre les deux bras de traitement.

Il est recommandé d'introduire un effet aléatoire patient et un effet aléatoire sur le temps afin de tenir compte des variations inter-sujets. Ces modèles nécessitent de préciser la structure de la matrice de variance-covariance entre les mesures de QdV (Littell *et al*, 2000). Un certain nombre de structure existent. Les quatre structures les plus souvent utilisées pour des données longitudinales sont : matrice non structurée (« unstructured », UN), autorégressive de premier ordre (« first-order autoregressive », AR(1)), autorégressive de premier ordre à hétérogénéité (« heterogeneous first-order autoregressive », ARH(1)), et hétérogène à symétrie composée (« heterogeneous compound symmetry », CSH). La matrice UN est la structure la plus générale. Elle est utilisée si aucune hypothèse ne peut être faite sur la structure de la matrice. Cependant, comme aucune contrainte n'est formulée, l'utilisation de cette matrice requiert l'estimation d'un grand nombre de paramètres. La structure ARH(1) considère que la corrélation entre les mesures de QdV décroit au cours du temps. La structure AR(1) suppose également que la corrélation entre les mesures de QdV décroit au cours du temps, mais aussi que les variances sont égales. Enfin, la structure CSH suppose que la corrélation entre les mesures de QdV reste constante au cours du temps. Généralement, ces quatre structures sont testées et le choix de la matrice se fait selon le critère Akaike Information Criteria (AIC).

Le modèle peut s'écrire :

$$Y_i = X_i\beta + Z_i\,\gamma_i + \varepsilon_i$$

Où :

- Y_i correspond au vecteur des scores du patient i ;
- β est un vecteur d'effets fixes ;
- Z_i est une sous-matrice de X_i pour les effets aléatoires ;
- γi est un vecteur d'effets aléatoires de distribution normale, de moyenne 0 et de matrice de covariance Σ ;
- et ε_i correspond à l'erreur aléatoire et suit une loi normale de moyenne 0 et d'écart type σ.

L'estimation des effets fixes et effets aléatoire peut être faite par la méthode du maximum de vraisemblance ou maximum de vraisemblance restreinte. Les effets fixes et effets aléatoires sont alors estimés l'un après l'autre. Les effets aléatoires sont estimés en premier (considérant

les effets fixes constants), puis les effets fixes (considérant les effets aléatoires constants) en maximisant la vraisemblance. Ces étapes sont itérées jusqu'à ce que la vraisemblance converge.

Une limite de ce modèle est qu'il requiert généralement la normalité du score. De plus, les résultats sont peu familiers pour les cliniciens.

Pour ces modèles, le calcul du nombre de sujets nécessaires est complexe est doit être réalisé sous un logiciel statistique adapté (Guo *et al*, 2013; Muller & Stewart, 2006). Ce calcul dépend:

- du taux d'erreur de type I,
- de l'hypothèse ciblée à tester,
- de la différence minimale de moyenne du score,
- des facteurs prédictifs : les catégories de chaque facteur prédictif doivent être précisées ;
- des variances attendues du score,
- et de la corrélation entre chaque paire de mesures répétées.

Ainsi, il peut être difficile de connaitre a priori la structure de la matrice de variance-covariance. Si celle-ci est mal choisie, la taille de l'échantillon peut ne pas être adaptée à l'analyse réalisée (Guo *et al*, 2013).

6.2.3. Generalized Estimating Equation

Les modèles « Generalized Estimating Equation » (GEE) sont utilisés pour estimés les paramètres d'un modèle linéaire généralisé avec une corrélation entre les mesures de QdV possiblement inconnue. Ces modèles ont déjà été appliqués dans plusieurs études de QdV (Coen *et al*, 2012; Pardo *et al*, 2010; Shi *et al*, 2011) mais restent peu exploités.

Ces modèles se focalisent sur la réponse moyenne et relient cette réponse marginale $\mu_{ij} = E(y_{ij})$ à une combinaison linéaire de covariables (Liang & Zeger, 1986; Zeger *et al*, 1988). Le modèle s'écrit ainsi :

$$g(y_{ij}) = x'_{ij}\beta$$

Avec :

- y_{ij} le score de QdV du patient i au temps j ;

- x'_{ij} le vecteur des co-variables explicatives ;
- β le vecteur des paramètres de régression ;
- $g(.)$ la fonction de lien.

Pour des réponses normalement distribuées, la fonction de lien est la fonction identité. Cependant, d'autres fonctions de lien peuvent être considérées telles que la fonction logit pour des réponses dichotomiques (de type « oui » vs. « non »).

Différentes matrices de covariances peuvent être choisies pour modéliser la corrélation entre les données. Les structures les plus souvent utilisées pour des données longitudinales sont : matrice non structurée, autorégressive de premier ordre, et interchangeable (supposant toutes les corrélations égales).

Dans ces modèles marginaux, la régression et la corrélation sont modélisées séparément, et l'une peut être modifiée sans nécessairement modifier la seconde (Diggle *et al*, 2002). L'estimation des paramètres se fait par une approche de quasi-vraisemblance (Wedderburn, 1974). Puisque les paramètres précisant la structure de la matrice de corrélation sont des paramètres de nuisances (ne sont pas des paramètres d'intérêt), des structures de matrices de corrélation simples telles que la structure interchangeable ou autorégressive de premier ordre peuvent être retenues pour préciser la corrélation intra-sujet. Les paramètres d'intérêt sont toujours valides même si la structure de la matrice de corrélation n'est pas correctement précisée (Liang & Zeger, 1986).

Jung *et al.* ont proposé une formule mathématique pour déterminer la taille d'échantillon nécessaire pour mettre en évidence un effet traitement au cours du temps entre deux bras de traitement (Jung & Ahn, 2003). Les auteurs ont également proposées quelques recommandations pour tenir compte de l'occurrence de données manquantes dans le calcul du nombre de sujets nécessaires.

6.2.4. Growth curve modeling

Les « Growth curve models » (ou modèles à courbe de croissance) sont utilisés notamment lorsque les temps d'évaluations diffèrent fortement entre les individus (Zee, 1998). Ils sont également pertinents lorsque le patient présente successivement plusieurs états de santé. Par

exemple, si l'on souhaite modéliser une évolution de la QdV au cours du traitement puis lors de la période de suivi (après arrêt du traitement).

Dans ces modèles, le temps est considéré comme une variable continue. Deux modèles sont généralement appliqués : un modèle polynomial ou un modèle de régression linéaire par morceaux.

Une fonction polynomiale est souvent utilisée pour modéliser l'évolution de la QdV (Fairclough, 2010). La forme générale du modèle est la suivante :

$$Y_i(t) = \beta_0 + \sum_k \beta_k t_i^k + \varepsilon_i$$

Avec

- $Y_i(t)$ le score de QdV du sujet i au temps t,
- β_i les coefficients du modèle,
- ϵ_i le terme d'erreur aléatoire.

Les termes quadratiques et cubiques permettent de s'écarter de la linéarité. Le nombre de paramètres (β_i) ne peut excéder le nombre de temps de mesure. Pour une étude avec 5 temps de mesure, le terme le plus élevé ne peut être supérieure à t_4.

Une limite de l'utilisation de ces modèles est leur interprétation. En effet, des termes cubiques ou quadratiques peuvent être difficiles à interpréter cliniquement.

Un modèle de régression linéaire par morceaux peut également être utilisé pour modéliser l'évolution de la QdV. Ce modèle permet donc de considérer une évolution de la QdV linéaire sur des intervalles de temps relativement courts. La forme générale du modèle est la suivante :

$$Y_i(t) = \beta_0 + \beta_1 t_i + \sum_{k \geq 2} \beta_k t_i^{[k]} + \varepsilon_i$$

$$t_i^{[k]} = \max(t_i - T^{[c]}, 0)$$

Ainsi, la trajectoire du modèle est $\beta_0 + \beta_1 t_i$ et le modèle change de trajectoire à $T^{[c]}$: la nouvelle trajectoire est $\beta_0 + \beta_1 t_i + \beta_2 t_i^{[k]}$. Comme pour le modèle polynomial, le nombre maximum de termes est égal au nombre de mesures de la QdV.

Les points où la trajectoire est modifiée doivent correspondre à des temps de mesure où la QdV est susceptible d'être modifiée comme à la suite du traitement par exemple.

6.2.5. Modélisation conjointe de données de QdV longitudinales et de survie

Une analyse conjointe des données de QdV longitudinales et des données de survies (données de type temps jusqu'à événement) peut être réalisée pour tenir compte de l'association et de la dépendance entre ces deux types de données (Ibrahim *et al*, 2010). Les données de survies peuvent correspondre à la survie globale ou la survie sans progression par exemple.
Dans ces modèles, les données de QdV sont généralement analysées par un modèle linéaire à effets mixtes tandis que les données de survies sont analysées par un modèle de Cox (Henderson *et al*, 2000). Ensuite, les composantes longitudinales et de survie sont liées par les coefficients issus du modèle mixte pour les données de QdV longitudinales :

$$Y_{ij} = X_{ij} + \epsilon_{ij} = \mu_0 + \mu_1 t + \gamma Z_i$$

$$h(t) = h_0(t)\exp(\beta X_{ij} + \alpha Z_i)$$

où

- Y_{ij} est le score de QdV du patient i au temps j ;
- X_{ij} correspond à la trajectoire du patient i au temps j ;
- ϵ_{ij} est le terme d'erreur aléatoire suivant une loi normale de moyenne 0 et d'écart type σ ;
- μ_0 est l'effet aléatoire patient ;
- μ_1 est l'effet aléatoire temps ;
- Z_i correspond à la variable indicatrice du bras traitement du patient i ;
- γ est l'effet du traitement ;
- α est l'effet direct du traitement sur la donnée de survie ;
- β mesure l'association entre les données de QdV longitudinales et la donnée de survie.

Trois types d'effets sont ainsi estimés :

- γ correspondant à l'effet du traitement sur la QdV ;
- α correspondant à l'effet du traitement sur la survie ;
- $\beta\gamma + \alpha$ correspondant à l'effet global du traitement.

Une analyse des données de QdV selon un modèle IRT a également été proposée (Wang *et al*, 2002) mais cela reste encore peu exploité.

Un calcul de la taille d'échantillon pour déterminer l'effet global du traitement ($\beta\gamma + \alpha$) a été proposé par Chen et al. (Chen *et al*, 2011). Les auteurs ont montré que la formule de Schoenfeld pouvait être étendue au design de l'étude de modélisation conjointe. Ce calcul est

assez simple et dépend uniquement de l'effet global du traitement ($\beta\gamma + \alpha$) et de la proportion de patients affiliés au bras de traitement 1 ($Z_i = 1$), en plus du taux d'erreur de type I et de la puissance souhaitée.

L'utilisation de modèles conjoints permet une augmentation de la puissance pour détecter l'effet globale dû au traitement $\beta\gamma + \alpha$ ainsi que l'effet du traitement sur la survie α.

6.2.6. QALYs et Q-TWIST

Une analyse de l'évolution de la QdV peut également être réalisée en tenant compte de différents états de santé exprimés par le patient depuis son diagnostic (Fayers & Machin, 2007). L'idée générale est de considérer que l'intervalle de temps entre le diagnostic du patient et sa survie globale peut être partitionné en différentes périodes distinctes durant lesquelles le niveau de QdV du patient peut changer. Il ne s'agit pas d'une analyse longitudinale a proprement dit de la QdV, mais plutôt d'une modélisation conjointe entre l'évolution de l'état de santé du patient et l'évolution de la QdV durant ces différents états. Cette approche permet de pondérer les bénéfices potentiels de la stratégie thérapeutique avec son impact sur la QdV du patient.

- **Préférence et utilité**

Les patients pourraient exprimer leur préférence pour un traitement vis-à-vis d'un autre traitement s'ils connaissaient les conséquences des deux stratégies sur leur QdV. Généralement, l'impact des traitements sur la QdV du patient n'est connu qu'avec une certaine probabilité d'incertitude. Lorsque les préférences du patient sont évaluées dans une telle situation où les conséquences du traitement ne sont pas certaines, on parle de critères d' « utilités ». Ces coefficients d'utilités peuvent être obtenus de différentes façons. Le plus souvent, ils proviennent de questionnaires de QdV (d'utilité) tels que :

- une échelle visuelle analogique où le patient évalue sa QdV actuelle et sa QdV souhaitée/probable selon divers scénarios
- le questionnaire EUROQoL EQ-5D évaluant 5 dimensions de QdV et conduisant à 243 états de santé possible.

Une fois que les coefficients d'utilités sont fixés, l'objectif est de les combiner avec le temps probable que le patient va passer dans chaque état de santé.

- **QALYs**

La méthode « Quality-Adjusted Life Years » (QALYs) ou « année de vie ajustée sur la qualité » permet de calculer un score global pour chaque patient tenant compte du temps passé dans les différents états de santé. Cette analyse ne doit pas être confondue avec une analyse de survie sans détérioration de la QdV.

Si trois états de santé sont considérés, alors le score QALY résumé du patient est :

$$QALY = U_1T_1 + U_2T_2 + U_3T_3$$

où U_i est le coefficient d'utilité de l'état i et T_i est le temps passé dans l'état de santé i.

Les valeurs de QALYs ainsi obtenues peuvent être comparées entre différents groupes de traitement. Cependant, cette comparaison repose sur certaines hypothèses :

- l'indépendance des coefficients d'utilité, i.e. qu'un individu évalue la quantité de vie dans un certain état de santé indépendamment de la valeur de cet état de santé ;
- la neutralité du risque, i.e. que les coefficients d'utilité sont une fonction linéaire des années de vie ;
- un compromis proportionnel constant, i.e. que la proportion de vie restante qu'un individu est prêt à sacrifier pour une amélioration de sa QdV est indépendante du nombre d'années de vie restantes.

Ces hypothèses sont rarement vérifiées et semblent peu réalistes.

Ces QALYS peuvent être utilisés pour une évaluation économique du traitement, il s'agit d'une évaluation « coût-utilité » (Goldhirsch *et al*, 1989). Le ratio « coût-utilité » de chaque traitement est alors calculé. La comparaison entre deux traitements A et B peut se faire en calculant le ratio de la différence de coût entre les deux traitements sur la différence de score QALYs.

- **Q-TWIST**

L'approche Q-TWIST (« Quality-adjusted Time WIthout Symptoms and Toxicity) est proche du concept des QALY-S. Cependant, contrairement aux QALYs, le Q-TWIST peut être appliqué aux données de survie contenant des données censurées (Fayers & Machin, 2007). L'idée générale des Q-TWIST est de partitionner les courbes de survies globales en plusieurs

régions définissant le temps passé dans des états cliniques particuliers. En général, trois états de santé sont considérés (Sloan *et al*, 2002) :

- L'intervalle de temps sans symptôme ni toxicité (« time without symptom or toxicity », TWIST)
- L'intervalle de temps passé avec des symptômes et toxicités (TOX)
- Et la progression de la maladie ou la rechute (PROG).

Le choix des différents états peut cependant varier selon les situations thérapeutiques.

Une fois que c'est états sont définies, l'idée est la même que pour les QALYs : résumer l'information par patient en une quantité unique en tenant compte des coefficients d'utilité de chaque état de santé :

$$Q - TWIST = U_{TWIST}TWIST + U_{TOX}TOX + U_{PROG}PROG$$

où U_{TWIST}, U_{TOX} et U_{PROG}sont les coefficients d'utilité de chaque état considéré.

Cette unique Q-TWIST est dans la même unité de temps que chaque mesure de temps considérée. Les moyennes de Q-TWIST par bras de traitement peuvent alors être calculées et comparées selon un test non paramétrique de Mann et Whitney. Les courbes de survies partitionnées selon les trois états de santé peuvent également être représentées (Sloan *et al*, 2002). Les Q-TWIST reposent également sur certaines hypothèses :

- Le temps ajusté sur la qualité passé dans un certain état de santé est proportionnel au temps réel passé dans cet état de santé ;
- la valeur du coefficient d'utilité d'un état de santé donné est la même quel que soit le (ou les temps) temps où l'état de santé est exprimé.

Une formule mathématique a été proposée pour déterminer la taille d'échantillon nécessaire pour designer des essais cliniques selon la méthode Q-TWIST en se basant sur la formule de la variance asymptomatique (Murray & Cole, 2000).

6.2.7. Méthode du temps jusqu'à détérioration d'un score de QdV

Depuis quelques années, une méthode d'analyse de survie basée sur le temps jusqu'à détérioration (TDJ) d'un score de QdV est régulièrement utilisée pour l'analyse longitudinale de la QdV, en particulier dans les essais cliniques de phase III (Bonnetain *et al*, 2010; Burris *et al*, 2013; Gourgou-Bourgade *et al*, 2013; Kabbinavar *et al*, 2008). Cette méthode a été proposée pour la première fois en 2002 par Award et al. (Awad *et al*, 2002).

Le TJD nécessite une définition de l'évènement, i.e. de la détérioration. Cette définition dépend du choix du score de référence, de la DMCI considérée, de la considération des scores manquants et de l'intégration ou non du décès dans la définition de l'évènement.

La définition la plus intuitive du TJD est le délai depuis la date d'inclusion ou de randomisation jusqu'à l'apparition d'une première détérioration significative du niveau de QdV par rapport au niveau à l'inclusion (Hamidou *et al*, 2011). Les patients ne présentant pas de détérioration sont censurés à la date de dernière mesure de la QdV ou de dernière nouvelle. Dans cette définition, la détérioration observée est un état transitoire et non nécessairement définitif.

Le temps jusqu'à détérioration définitive (TJDD) d'un score de QdV a été défini en situation avancée comme l'intervalle de temps depuis la date d'inclusion ou de randomisation et l'apparition d'une première détérioration significative par rapport au score obtenu à l'inclusion, sans amélioration ultérieure significative par rapport au score à l'inclusion, ou le décès (Bonnetain *et al*, 2010). Un patient présentant une détérioration significative du niveau de QdV suivi de données manquantes monotones dû à la sortie de l'étude du patient (drop-out) ou au décès est considéré en détérioration définitive. Les patients ne présentant pas de détérioration et ceux présentant une détérioration mais suivie d'une amélioration significative par rapport au score à l'inclusion sont censurés au moment du remplissage du dernier questionnaire de QdV.

Les données manquantes intermittentes ne sont généralement pas prises en compte dans ces modèles, considérant que le niveau de QdV reste constant entre deux mesures de QdV consécutives.

Pour les questionnaires de l'EORTC, la DMCI pour caractériser la détérioration est en général de 5 ou 10 points (Bonnetain *et al*, 2010; Gourgou-Bourgade *et al*, 2013; Hamidou *et al*, 2011). Pour les questionnaires du groupe FACT, la DMCI serait de 5 à 7 points du score global (Wu *et al*, 2011).

Une détérioration d'au moins un score de QdV parmi un ensemble de scores peut également être considérée (Bonnetain *et al*, 2010; Hamidou *et al*, 2011). Dans ce cas, la date de détérioration retenue est celle de la première détérioration observée, quel que soit le score affecté.

S'agissant d'une méthode d'analyse de survie, son estimation se fait généralement par la méthode d'estimation de Kaplan-Meier (Goel *et al*, 2010). La courbe de survie selon la méthode de Kaplan-Meier est définie par la probabilité de ne pas présenter l'évènement considéré à un temps de mesure donné considérant le temps comme plusieurs intervalles réguliers. Cette méthode est basée sur l'idée intuitive que d'être en vie au temps T nécessite d'être en vie juste avant le temps T et de ne pas décéder au temps T. Dans cette définition, l'évènement est le décès. Pour la méthode du TJD/TJDD, l'événement est la détérioration, définitive ou non, d'un ou plusieurs scores de QdV. La méthode d'estimation de Kaplan-Meier correspond à la formule suivante:

$$S(t) = \prod_{t_i \leq t} \frac{n_i - m_i}{n_i}$$

où $n_i = n_{i-1} - m_{i-1} - c_{i-1}$

et:

- n_i est le nombre de patients à risque au temps i, i.e. ne présentant pas de détérioration est encore présent dans l'étude ;
- m_i est le nombre de détériorations observées au temps i ;
- et c_i correspond au nombre de patients censurés au temps $i - 1$.

Les résultats sont ensuite généralement comparés par bras de traitement à l'aide d'un test de Log-Rank et résumé par un Hazard Ratio obtenu par un modèle de Cox univarié (Cox, 1972).

Cette méthode d'analyse est très attirante pour les cliniciens puisqu'elle propose des résultats facilement compréhensibles et interprétables, résumés par des Hazard Ratio et des médianes de TJD par bras de traitement. De plus, les résultats sont souvent en cohérence avec ceux obtenus pour le critère de jugement principal, soit la survie globale ou la survie sans progression (Burris *et al*, 2013; Gourgou-Bourgade *et al*, 2013).

Par ailleurs, une taille d'échantillon nécessaire peut être facilement obtenue pour la méthode du TJD d'après la formule de Schoenfeld (Schoenfeld, 1983). Ainsi, le nombre de sujets nécessaire dépend de :

- la proportion attendue π_A de patients détériorés dans le bras A au temps T ;
- la proportion attendue π_B de patients détériorés dans le bras B au temps T ;
- la taille Δ de l'effet attendu, soit l'Hazard Ratio ;
- le taux d'erreur de type I ;

- et la puissance statistique souhaitée.

La méthode du TJD présente cependant certaines difficultés méthodologiques. En effet, la méthode d'estimation de Kaplan-Meier pourrait ne pas être la plus adaptée aux évaluations de QdV. L'intervalle de temps entre les mesures de QdV peut influencer l'estimation de Kaplan-Meier et ainsi entrainer une surestimation du TJD/TJDD comme cela a déjà été démontré pour la survie sans progression (Panageas *et al*, 2007). Le « vrai » temps de détérioration du score de QdV étant a priori inconnu (sauf si l'évaluation de la QdV est faite quotidiennement), des approches statistiques adaptées aux évaluations par intervalle pourraient donc être proposées (Gong & Fang, 2013; He *et al*, 2013).

Des travaux ont déjà été menés sur la méthodologie de la méthode du temps jusqu'à détérioration et en particulier sur l'utilisation du test du Log-rank pour un taux de détérioration fixe ou variable (Boisson, 2008).

L'utilisation d'un modèle de régression de Cox pour l'estimation de la différence entre les bras de traitement et pour l'analyse multivariée repose par ailleurs sur l'hypothèse de proportionnalité des risques (Cox, 1972). Si cette hypothèse n'est pas respectée, différentes modélisations peuvent être proposées dont le temps de survie moyen restreint (« restricted mean survival time », RMST) (Royston & Parmar, 2011; Royston & Parmar, 2013). Le RMST correspond à l'aire sous la courbe de survie entre t=0 et t=t* (t restreint). Ainsi, pour deux bras de traitement A et B de courbe de survie respectives $S_0(t)$ et $S_1(t)$, la différence Δ de RMST entre les deux bras est :

$$\Delta = \int_0^{t^*} S_1(t) - \int_0^{t^*} S_0(t)$$

et correspond à l'aire entre les deux courbes de survie.

Enfin, la méthode du TJD pourrait également rencontrer des problèmes de risques compétitifs, que ce soit entre les scores des différentes dimensions de QdV, ou entre les scores de QdV et le décès. L'occurrence éventuelle de ces risques compétitifs doit être étudiée par l'analyse des courbes d'incidence cumulées des évènements potentiellement en compétition. Les courbes d'incidence cumulées sont alors comparées par un test de Gray (Gray, 1988). Un modèle multivarié de Fine et Gray peut ensuite être construit (Fine & Gray, 1999).

6.2.8. ***Modèles IRT pour l'analyse longitudinale***

Les modèles IRT ont d'ores et déjà montré leur intérêt dans la validation des propriétés psychométriques des questionnaires. Ils pourraient également être appropriés pour l'analyse longitudinale de données de QdV. Depuis quelques années, ils commencent ainsi à être étendus à l'analyse longitudinale (Douglas, 1999; Glas *et al*, 2009).

- **Modèle de Rasch longitudinal**

Le modèle de Rasch a ainsi été étendu à l'analyse longitudinale par Meiser en 2007 (Meiser, 2007).

Dans le modèle de Rasch, la probabilité p_{ij} qu'a l'individu i de choisir la réponse x (= 0 ou 1) à l'item j au temps t connaissant son niveau de trait latent θ_i et la difficulté δ_j de l'item j est :

$$p_{ij} = P(X_{ij} = x|\theta_i, \delta_j) = \frac{\exp(\theta_i - \delta_j)}{1 + \exp(\theta_i - \delta_j)}$$

Dans le modèle de Rasch longitudinal, la probabilité p_{ij} qu'a l'individu i de répondre positivement à l'item j connaissant son niveau de trait latent au temps t ${\theta_i}^{(t)}$et la difficulté δ_j de l'item j est :

$$p_{ijt} = P({X_{ij}}^{(t)} = x^{(t)}|{\theta_i}^{(t)}, \delta_j) = \frac{\exp({\theta_i}^{(t)} - \delta_j)}{1 + \exp({\theta_i}^{(t)} - \delta_j)}$$

Les paramètres de difficulté des items restent constants au cours du temps. La valeur du trait latent pour l'individu i $\theta_i = ({\theta_i}^{(1)}, \dots, {\theta_i}^{(t)})$ est supposée suivre une loi normale multivariée de moyenne 0 et de matrice de variance covariance Σ. Ce modèle ne peut s'appliquer qu'aux items dichotomiques, peu utilisés dans les questionnaires de QdV. Une extension du modèle PCM à l'analyse longitudinale pourrait être envisagée de la même façon.

On peut alors modéliser l'évolution du trait latent pour le patient i ${\theta_i}^{(t)}$ au cours du temps par des effets fixes et effets aléatoires, de la même façon que les modèles linéaires à effets mixtes basés sur le score. A titre d'illustration, l'équation suivante correspond à la modélisation de l'évolution du trait latent avec des effets fixes bras de traitement, temps et temps interaction

traitement et des effets aléatoires patients et temps pour refléter les déviances individuelles des patients à l'inclusion et au cours du temps :

$$\theta_i{}^{(t)} = a * arm_i + b * t + \ c * arm_i * t + \ + u_{0,i} + u_{1,i} * t$$

$$\left(u_{0,i}, u_{1,i}\right) \sim N\left(\binom{0}{0}, \Sigma\right)$$

$$\varepsilon_i^{(t)} \ \sim N(0, \sigma^2)$$

avec:

- $bras_i$ est le bras de traitement du patient i (égal à 0 ou 1) ;
- a l'effet fixe bras de traitement ;
- b l'effet fixe temps ;
- c l'effet fixe d'interaction entre le temps et le bras de traitement ;
- $u_{0,n}$ l'effet aléatoire patient ;
- $u_{1,n}$ l'effet aléatoire temps ;
- $\varepsilon_i^{(t)}$ le terme d'erreur résiduelle du patient i au temps t ;
- σ^2 la variance résiduelle.

- **Linear Logistic Model with Relaxed Assumptions**

Une autre approche a été proposée par Fisher en 1995 (Fisher, 1995). Il s'agit du modèle linéaire logistique à hypothèses relâchées (« Linear Logistic Model with Relaxed Assumptions », LLRA). Contrairement au modèle de la famille de Rasch, ce modèle ne requiert pas d'unidimensionnalité du trait latent et ne fait aucune hypothèse de distribution de la population étudiée au regard du trait latent (d'où son nom de modèle à hypothèses relâchées). De plus, ce modèle est adapté aux items polytomiques et a été développé pour mesurer le changement entre plusieurs temps d'évaluation.

L'idée principale de ce modèle est de considérer, non pas un changement longitudinal par un changement au niveau des paramètres des individus (paramètre de capacité des individus), mais par un changement au niveau des paramètres des items (i.e. des paramètres de difficulté des items). Ainsi, les paramètres de capacités des individus sont fixés au cours du temps et seuls les paramètres des items varient. Puisque les paramètres de capacité des individus peuvent être considérés comme des paramètres de nuisances et sont généralement plus nombreux que les paramètres d'items, les paramètres de changement au niveau des items

peuvent être estimés au lieu des paramètres de changement au niveau des capacités des individus par maximum de vraisemblance conditionnel (Mair & Hatzinger, 2007a).

Dans ce modèle, un item j de paramètre de difficulté δ_j pour lequel un individu répond à deux temps de mesure peut être vu comme une paire d'items virtuels j_1^* et j_2^* de paramètres respectifs $\delta_{j_1}^*$ et $\delta_{j_2}^*$. Pour le premier temps de mesure, $\delta_{j_1}^* = \delta_j$ alors que pour le second temps de mesure, $\delta_{j_2}^* = \delta_j + \tau$ où τ est le paramètre de tendance (changement) du paramètre de l'item. C'est ce paramètre de tendance qui est ciblé par le modèle LLRA.

Pour les items polytomiques, le modèle se construit soit à partir du modèle PCM si on considère un nombre variable de modalités de réponse par items ou des distances entre les catégories de réponse différentes d'un item à un autre, soit à partir du RSM dans le cas contraire.

Ainsi, selon l'approche à crédit partiel (PCM), le modèle peut s'écrire :

$$\mathrm{P}\left(X_{ijk} = 1 \middle| T_1\right) = \frac{\exp(k\theta_{ij} + \omega_{jk})}{\sum_{l=0}^{m_j} \exp(l\theta_{ij} + \omega_{jl})}$$

$$\mathrm{P}\left(X_{ijk} = 1 \middle| T_2\right) = \frac{\exp(k\theta'_{ij} + \omega_{jk})}{\sum_{l=0}^{m_j} \exp(l\theta'_{ij} + \omega_{jl})} = \frac{\exp(k\left(\theta_{ij} + \beta_{ij}\right) + \omega_{jk})}{\sum_{l=0}^{m_j} \exp(l\left(\theta_{ij} + \beta_{ij}\right) + \omega_{jl})}$$

avec:

- θ_{ij} paramètre de capacité de l'individu i au niveau du j-ème trait latent au temps T_1 ;
- θ'_{ij} paramètre de capacité de l'individu i au niveau du j-ème trait latent au temps T_2 ;
- β_{ij} changement de capacité de l'individu i au niveau du trait latent j ;
- ω_{jk} paramètre de la modalité de réponse k pour l'item j ;
- $m_j + 1$ nombre de modalités de réponse pour l'item j.

L'idée générale de ce modèle est de décomposer le paramètre de changement β_{ij} de l'item en une combinaison linéaire d'effets $\beta_{ij} = W_j^T \eta$ où W_j^T représentent les valeurs des covariables pour le trait latent j (poids) et η est un vecteur de paramètres dits de base pouvant contenir les paramètres des items ainsi que des effets temps et des effets groupes.

Ce modèle a été implémenté sous le logiciel R sous la forme d'un package eRM (extended Rasch Modeling) (Mair & Hatzinger, 2007b).

Les modèles IRT pour l'analyse longitudinale restent encore peu exploités. A ce jour, ils sont généralement explorés via des données de QdV simulées (Blanchin *et al*, 2011a; Blanchin *et al*, 2011b; de Bock *et al*, 2013; de Bock *et al*, 2014) et sont peu appliqués dans le cadre d'essais cliniques en cancérologie (Douglas, 1999; Glas *et al*, 2009). Il existe donc peu d'information sur la capacité de ces modèles à mettre en évidence un effet traitement. Les études de simulations déjà réalisées se sont focalisés sur des études avec trois temps de mesure où la QdV est évaluées avec des items dichotomiques (Blanchin *et al*, 2011a; Blanchin *et al*, 2011b; de Bock *et al*, 2013; de Bock *et al*, 2014), ce qui est peu utilisés dans les questionnaires de QdV en cancérologie. De plus, ces modèles n'ont pas encore été appliqués pour une analyse longitudinale multivariée.

Les premières recommandations pour une taille d'échantillon dans le cadre d'une étude transversale de comparaison entre deux groupes ont récemment été données (Blanchin *et al*, 2013; Sebille *et al*, 2014). Cependant, la détermination d'une taille d'échantillon selon les modèles longitudinaux est encore en cours d'investigation.

Il parait également nécessaire d'implémenter ces modèles d'IRT pour l'analyse longitudinale sous différents logiciels statistiques pour pouvoir utiliser facilement ces modèles. Enfin, ces modèles restent particulièrement complexes et les résultats proposés peuvent paraître difficile à aborder pour les cliniciens.

6.3. Problématique des données manquantes

Une difficulté majeure de l'analyse longitudinale de la QdV est l'occurrence de données manquantes (Fairclough, 2010; Fairclough *et al*, 1998). En effet, il est rare que les questionnaires soient complétés intégralement par le patient et ce à chaque temps de mesure. Selon les situations, les données manquantes observées peuvent être informatives ou non de l'état de santé ou de QdV du patient. Si les données manquantes sont dépendantes de l'état de santé ou de la QdV du patient, le risque est de travailler sur une sous-population de patient en meilleur état de santé, ou bien de surestimer par exemple l'état de santé du patient. Il est donc

nécessaire d'étudier le profil des données manquantes afin de pouvoir en tenir compte de façon adéquate dans la méthode d'imputation des scores et/ou dans l'analyse longitudinale (Fairclough, 2010).

6.3.1. Définition et types de données manquantes

Le terme de donnée manquante désigne diverses situations :

- Des données manquantes prévues ou non ;
- Des données manquantes à l'échelle de l'item ou du questionnaire ;
- Des données manquantes intermittentes ou monotones.

Dans les essais cliniques en cancérologie, la QdV est généralement mesurée jusqu'à la sortie d'étude du patient (« drop-out ») pouvant correspondre à la date de progression de la tumeur ou au décès du patient. Dans ce cas, aucune donnée de QdV n'est disponible après la sortie d'étude du patient. Des données manquantes sont donc observées suite à la sortie d'étude du patient. Ce phénomène est appelé **attrition**. Ainsi, la donnée manquante observée est prévue.

La **non-compliance** au questionnaire correspond à l'absence de remplissage du questionnaire à un temps de mesure donné alors que le patient devait remplir le questionnaire. Il s'agit donc ici d'une donnée manquante non prévue.

Ainsi, il est nécessaire de donner les différents taux de remplissage des questionnaires à chaque temps de mesure :

- Taux de questionnaires manquants ou totalement vides ;
- Taux de questionnaires complètement et partiellement rempli :
    - Par rapport à l'ensemble des patients ;
    - Par rapport à l'ensemble des patients encore présents dans l'étude au temps théorique de remplissage (correspond au taux de compliance) ;
- Taux d'items manquants par individu et par questionnaire.

Les données manquantes observées peuvent être :

- **intermittentes**, i.e. la donnée manquante a lieu à un ou plusieurs temps de mesure donnés, mais des données de QdV sont toujours disponibles à un temps de mesure ultérieure ;
- **monotones**, i.e. toutes les données de QdV sont manquantes à partir d'un certain temps T.

Une donnée manquante intermittente peut correspondre à un item manquant parmi l'ensemble du questionnaire ou de la dimension concernée, on parle alors d'item manquant intermittent (« **intermittent missing item** ») (Fayers *et al*, 1998). Il se peut également qu'un patient n'a pas pu répondre à un questionnaire à un temps de mesure donné, indépendamment ou non de son état de santé. Il s'agit alors d'un questionnaire manquant de façon intermittente (« **intermittent missing form** »), à partir du moment ou des données de QdV sont observées ultérieurement (Curran *et al*, 1998b). Enfin, lorsqu'un patient quitte l'étude prématurément, généralement dû à une détérioration de son état de santé, il s'agit d'un phénomène de « drop-out » et il en résulte des données manquantes monotones (Diggle & Kenward, 1994). Si le patient décède au cours de l'étude, des données manquantes monotones suite au décès seront également observées.

6.3.2. Classification des données manquantes

Little et Rubin ont proposés une classification des données manquantes selon trois profils (Little & Rubin, 1987) : données manquantes complètement aléatoires, données manquantes aléatoires et données manquantes non aléatoires (Tableau 5).

Les données manquantes dues au hasard, i.e. indépendantes de l'état de santé du patient et des données observées, comme les données cliniques et socio-démographiques recueillies à l'inclusion, sont considérées comme des données manquantes complètement aléatoires (« **Missing Completely At Random** », MCAR). Par exemple, si un patient a oublié de remplir un questionnaire à un temps de mesure donné ou certains items parmi un questionnaire, alors les données manquantes observées sont complètement dues au hasard.

Les données manquantes sont dites manquantes aléatoirement (« **Missing At Random** », MAR) si elles dépendant de données observées mais sont indépendantes des données non observé (donc ici du niveau de QdV manquant du patient). A titre d'exemple, les femmes ou les personnes âgées peuvent trouver certaines questions relatives à leur sexualité dérangeantes et ne pas souhaiter répondre à ces questions. Ces données manquantes peuvent donc être expliquées par des données sociodémographiques recueillies à l'inclusion. Elles peuvent être également expliquées par le niveau de QdV observé à l'inclusion par exemple. Conditionnellement à ces données observées, ces données manquantes seront considérées comme MCAR.

Enfin, les données manquantes sont considérées comme manquantes non aléatoirement (« **Missing Not At Random** », MNAR) si elles dépendent du niveau de QdV non observé du patient. Par exemple, si un patient ne peut pas remplir un questionnaire à un temps d'évaluation donné en raison d'une hospitalisation passagère ou d'un niveau de fatigue trop important ; ou si un patient quitte prématurément l'étude en raison d'une détérioration de son état de santé, alors les données manquantes observées dépendent du niveau de QdV non observé du patient et sont de type MNAR.

Tableau 5 : Type de données manquantes pour la QdV selon la classification de Little et Rubin (Little & Rubin, 1987)

Données manquantes	Dépend de	Ne dépend pas de
Complètement aléatoires (MCAR)		Variables observées Niveau de QdV non observé
Aléatoire (MAR)	Covariables observées Niveau de QdV observé	Niveau de QdV non observé
Non aléatoire (MNAR)	Niveau de QdV non observé +/- variables observées	

6.3.3. Impact des données manquantes

Les données manquantes de type MCAR et MAR sont non informatives de l'état de santé et du niveau de QdV du patient. Elles n'entraineront donc pas de biais dans l'analyse longitudinale de la QdV si elles sont correctement traitées. En effet, pour les données manquantes de type MAR, il est nécessaire de détecter les données observées dépendantes de la donnée manquantes. La donnée manquante pourra alors être considérée comme MCAR conditionnellement à ces données observées. En revanche, ces données manquantes de type MCAR et MAR peuvent entrainer une perte de puissance pour l'analyse longitudinale. Comme le nombre de données observées sera inférieur au nombre de données prévues initialement, il sera plus difficile de mettre en évidence une différence de QdV entre deux types de traitement, même si cette différence existe.

Les données manquantes de type MNAR, quant à elles, sont informatives de l'état de santé et de la QdV du patient. Elles peuvent entrainer à la fois une perte de puissance et un risque de biais dans l'analyse longitudinale si elles ne sont pas prises en compte de façon adéquate dans l'analyse (Fairclough *et al*, 1998). On distingue entre autre l'impact des sorties d'étude prématurées dues à une détérioration de l'état de santé ou au décès du patient (Diggle & Kenward, 1994). Ainsi, le nombre de données observées est inférieur au nombre de données

initialement prévues dans l'étude et seul les patients en meilleur état de santé restent dans l'étude.

En raison de l'impact potentiel des données manquantes, il est nécessaire de déterminer le profil des données manquantes afin d'en tenir compte de façon adéquate dans la méthodologie d'analyse des données de QdV.

6.3.4. Détermination du profil des données manquantes

Lorsque le pourcentage de données manquantes est très faible (jusqu'à 5%), on peut généralement considérer que les données manquantes sont MCAR sans risquer d'entrainer de biais dans l'analyse. Au-delà de ce pourcentage, il est important d'étudier le profil des données manquantes. L'identification du profil des données manquantes reste un processus complexe (Curran *et al*, 1998a).

L'identification d'un profil MAR de données manquantes vs. MCAR peut se faire en comparant les patients ayant complétés le questionnaire de QdV à l'inclusion aux autres patients. Selon les effectifs et les situations, les profils de patients comparés peuvent être :

- Les patients ayant un questionnaire complet à l'inclusion versus les autres patients ;
- Les patients ayant tous les scores estimables à l'inclusion versus les autres. Pour les questionnaires de l'EORTC et du groupe FACT, un score peut être calculé si au moins 50% des items de la dimension sont complétés.

Cette comparaison se fait selon les caractéristiques sociodémographiques et cliniques des patients recueillies à l'inclusion. Un modèle de régression logistique multivarié peut être utilisé afin de déterminer les variables indépendamment associées à l'occurrence de données manquantes. Néanmoins, des tests univariés sont le plus souvent utilisés. Un test du χ^2 ou un test exact de Fisher sera ainsi utilisé pour les variables catégorielle tandis qu'un test T de Student ou un test non paramétrique de Mann et Whitney pour les variables continues.

Cette analyse permettra de mettre en évidence des profils de données manquantes aléatoires (MAR), i.e. dépendant de covariables caractéristiques des patients.

Pour tester si les données manquantes dépendent du niveau de QdV observé du patient, une approche graphique peut être réalisée. Il s'agit ici de représenter le niveau moyen de QdV observé du patient défini par leur schéma (« pattern ») de données manquantes, par exemple

en stratifiant par rapport au dernier temps disponible. Ainsi, si les patients qui sortent de l'étude précocement présentent un niveau de QdV à l'inclusion plus faible que les autres patients ; et/ou une diminution du niveau de QdV juste avant la sortie d'étude, alors les données manquantes dépendent du niveau de QdV antérieur du patient et sont de type MAR.

Little a également proposé un test formel pour tester l'hypothèse de données manquantes de type MCAR vs. MAR (Little, 1988). L'idée de base de ce test est de comparer les moyennes observées pour chaque pattern de données manquantes. Si les données manquantes sont de type MCAR, alors les moyennes doivent être identiques pour chaque pattern.

La détermination du profile MNAR reste complexe puisque dans ce cas, la présence de donnée manquante est liée au niveau de QdV manquant du patient. En revanche, il est possible de recueillir certaines preuves d'un tel profil, en particulier lorsque d'autres variables indicatrices de la maladie ou du traitement recueillies sont fortement corrélées avec la QdV (Fairclough, 2010). Par exemple, si la donnée manquante est associée avec la progression de la maladie, alors il est vraisemblable que la donnée manquante soit de type MNAR. Recueillir les motifs de non remplissage des questionnaires permet également de déterminer si la donnée manquante est de type MNAR ou non. Ainsi, si le patient n'a pas rempli le questionnaire car il était trop fatigué, hospitalisé en lien avec son cancer, ou pour une autre raison, alors la donnée manquante est informative de l'état de santé et de la QdV du patient (donnée manquante MNAR). En revanche, si le patient n'a pas pu remplir le questionnaire car sa visite à l'hôpital a été déplacée ou parce qu'il a déménagé, alors la donnée manquante est indépendante de l'état de santé du patient et vraisemblablement de type MCAR.

6.3.5. Gestion des données manquantes

Selon le profil (MCAR/MAR/MNAR) et le type (intermittentes/monotones) de données manquantes, différentes stratégies d'analyses statistiques peuvent être proposées.

- **Complete case analysis**

La méthode la plus simple pour analyser des données en présence de données manquantes est de supprimer de l'analyse tous les patients présentant au moins une donnée manquante à un temps de mesure donné : il s'agit de l'analyse des cas complets (« complete case analysis »). Seuls les patients avec l'ensemble des données disponibles à chaque temps de mesure sont

donc analysés. Cette méthode d'analyse était très souvent utilisée par le passé, beaucoup moins de nos jours. Si les données manquantes sont de type MCAR, cette méthode entraîne uniquement une perte de puissance due au nombre réduit de sujets inclus dans l'analyse. Compte tenu du phénomène d'attrition, l'échantillon analysé peut être très petit par rapport à l'échantillon initial. Cette méthode est fortement déconseillée compte tenu de l'hypothèse forte de données manquantes MCAR et de la baisse de puissance importante (Fairclough, 2010).

- **Available case analysis**

Une autre méthode serait d'analyser les données disponibles (« available case analysis »). Ainsi, si l'on souhaite étudier la QdV à un temps donné, les patients pour lesquels le score de QdV est disponible à ce temps de mesure seront inclus dans l'analyse. Le principal inconvénient de cette méthode est que chaque analyse ne sera pas nécessairement menée sur le même échantillon de patient. Ainsi, l'interprétation des résultats risque d'être difficile. De plus, si les données manquantes sont de type MNAR, l'analyse menée sur les cas disponibles sera biaisée : l'analyse sera menée sur un sous-échantillon de patients présentant un meilleur état de santé.

- **Méthodes d'imputation des items**

Pour éviter les problèmes rencontrés avec l'analyse des cas complets ou des cas disponibles, une option fréquemment utilisée est de remplacer les données manquantes par des valeurs imputées en tenant compte des données observées.

Cette imputation peut concerner :

- les items manquants parmi une dimension donnée d'un questionnaire ;
- et/ou les questionnaires totalement manquants.

De plus, l'imputation est dite :

- simple si la donnée manquante est remplacée par une valeur unique ;
- multiple si la donnée manquantes est remplacée par plusieurs valeurs possibles.

Différentes méthodes d'imputations des items ont été proposées (Peyre *et al*, 2011).

- Personal Mean Score imputation

Lorsque des items sont manquants, la recommandation de la majorité des manuels de scoring des questionnaires est de considérer que l'item manquant ne diffère pas significativement des items renseignés, à condition qu'au moins un certain pourcentage d'items de la dimension considérée soient renseignés. Pour les questionnaires des groupes EORTC et FACT par exemple, le pourcentage requis est de 50%. Cette méthode d'imputation est appelée méthode d'imputation simple par la moyenne par individu (ou « personal mean score imputation»). Cette méthode est simple à réaliser, cependant elle requiert que les items manquants soient manquants de façon complètement aléatoire (MCAR). Cette méthode serait donc à éviter dans le cas de données manquantes qui ne serait pas complètement dues au hasard, en particulier si les données manquantes dépendent du niveau de la QdV non observé du patient (Fielding *et al,* 2008). En effet, le recours à cette méthode peut entrainer une surestimation ou une sous-estimation du niveau de QdV du patient en se focalisant uniquement sur les items renseignés. Cette méthode nécessite également que les items constituant l'échelle soient de niveaux de difficulté similaires, selon la terminologie utilisée dans les modèles IRT.

Néanmoins, dans la plupart des essais cliniques où la QdV a été mesurée comme critères de jugement secondaire ou tertiaire, la méthode d'imputation des scores est souvent passée sous silence de même que l'étude du profil des données manquantes, ce qui sous-entend une imputation simple par la moyenne.

- Item Mean Score imputation

Une méthode d'estimation par la moyenne par item peut également être réalisée (« item mean score imputation») (Huisman, 2000). La moyenne par item peut être calculée pour l'ensemble des patients ou par sous-groupes, en tenant compte de variables caractéristiques des individus afin de tenir compte éventuellement de données manquantes de type MAR. Contrairement à la méthode d'imputation simple par individu, cette méthode permet de tenir compte de la difficulté de l'item, selon la terminologie des modèles IRT. Cependant, elle ne tient pas compte du niveau de QdV propre au patient et reste peu utilisée en pratique.

- Modèle de régression

Un modèle de régression peut également être utilisé pour l'imputation des données (Chavance, 2004). Cette méthode consiste à construire un modèle de régression logistique ordinal où la variable à expliquer est l'item présentant des données manquantes et les variables dépendantes sont les autres items de la dimension. Des variables socio-

démographiques et cliniques peuvent également être ajoutées au modèle. L'estimation du modèle est réalisée d'après les données des patients ayant répondu à l'item. Les valeurs prédites obtenues par le modèle permettent ensuite d'estimer la valeur manquante pour l'item d'intérêt. Un terme d'erreur aléatoire peut être ajouté au modèle afin de refléter l'incertitude liée à la donnée manquante. L'avantage de cette méthode est qu'elle permet de tenir compte des données manquantes de type MAR, i.e. dépendant de variables caractéristiques des patients. Cette méthode est également simple à réaliser mais elle reste peu appliquée dans les analyses de QdV.

- Imputation selon un modèle IRT

Un modèle d'IRT peut également être utilisé pour l'imputation des items (Hardouin *et al*, 2011). Un modèle itératif peut alors être appliqué de la façon suivante :

- les données manquantes sont remplacées par les valeurs estimées par le modèle ;
- les paramètres du modèle d'IRT sont à nouveau estimés ;
- de nouvelles valeurs sont obtenues pour les données manquantes ;
- le processus s'arrête lorsque deux itérations successives donnent les mêmes valeurs pour les données manquantes.

Cette méthode permet de tenir compte à la fois du niveau de QdV de l'individu et de la difficulté de l'item. Néanmoins, elle nécessite que l'échelle ait un bon ajustement au modèle IRT et reste complexe à mettre en œuvre.

- Imputations multiples

Une méthode d'imputations multiples basée sur un algorithme de Monte Carlo par Chaîne de Markov peut également être utilisée (Cole *et al*, 2005). Cette méthode permet de tenir compte du caractère incertain de la vraie valeur de la donnée manquante. Cette méthode permettrait donc de ne pas sous-estimer la variance des données. Il serait recommandé de faire jusqu'à 5 imputations par donnée manquante (Fairclough, 2010). Cinq jeux de données sont alors créés et les analyses se font par jeu de données. Une procédure permet ensuite de combiner les résultats obtenus. Le logiciel de statistiques SAS ® (Version 9.3, SAS Institute Inc, Cary, NC) permet en particulier de faire ces analyses. La procédure MI permet de réaliser l'imputation multiple selon l'algorithme de Monte Carlo par Chaîne de Markov et la procédure MIANALYZE permet ensuite de combiner les résultats des diverses analyses. Cette méthode est particulièrement recommandée lorsque les données manquantes dépendent de données observées comme l'âge ou le sexe des patients. L'imputation peut alors se faire en tenant

compte de ces facteurs associés à l'occurrence de données manquantes. De plus, la procédure peut tenir compte des mesures de QdV antérieures. Il peut paraître difficile de mettre en œuvre une méthode d'imputation complexe comme des imputations multiples à chaque temps de mesure puisque le nombre de patients ayant un questionnaire à chaque temps de mesure diminue (phénomène d'attrition). L'algorithme aura par ailleurs des difficultés à converger si trop peu de données sont disponibles. Cette méthode est davantage utilisée pour l'imputation des questionnaires totalement vides.

Bien que de nombreuses méthodes aient été proposées, il n'existe pas à ce jour de standard pour l'imputation des items manquants. Il est de plus difficile de déceler des données manquantes de profil MNAR puisque celles-ci dépendraient du niveau de QdV non observé du patient. La méthode d'imputation « Personal Mean Score » reste à ce jour la procédure la plus simple et la plus utilisée.

Il existe également des méthodes d'imputations des questionnaires manquants :

- Last Observation Carried Forward

La méthode d'imputation « Last Observation Carried Forward » (LOCF) consiste à remplacer la donnée manquante observée au temps T par la dernière valeur observée précédemment. Ainsi, si un patient a rempli le questionnaire à l'inclusion, mais n'a pas rempli le questionnaire au temps de mesure suivant, alors on considère que le niveau de QdV du patient est resté constant au deuxième temps par rapport à la mesure à l'inclusion. Cette méthode est très simple à réaliser. Néanmoins, le principal inconvénient de cette méthode est de considérer le niveau de QdV du patient constant en présence de données manquantes. En effet, si la première mesure de QdV a lieu avant le début du traitement et que la seconde mesure (manquante) a lieu durant, le traitement, il est susceptible que le patient présente des effets secondaires dus au traitement qui peuvent impacter le niveau de QdV du patient. De plus, en cas de sortie d'étude du patient (« drop-out »), cette méthode émet l'hypothèse que le niveau de QdV du patient reste constant après la sortie d'étude ce qui semble inapproprié dans la plupart des études.

- Valeurs arbitraires de type « Worst case »

Une autre méthode d'imputation consiste à substituer la valeur manquante par une valeur arbitraire élevée ou fiable. Cette méthode est fréquemment utilisée lorsque la donnée manquante résulte d'un évènement indésirable grave tel que le décès du patient (Fairclough *et al*, 1999; Raboud *et al*, 1998). Ainsi, une valeur de QdV égale à 0 peut être utilisée en cas de décès. En cas de sortie d'étude en raison d'un niveau de toxicité élevé, une valeur de QdV égale au plus faible niveau de QdV observé peut être utilisée. Ces valeurs arbitraires permettent de tenir compte de données manquantes MNAR mais le choix des valeurs peut être très controversé et ne pas être adapté à toutes les situations de drop-out (Heyting *et al*, 1992).

- Modèles de régression

Comme pour les items manquants, des modèles de régression peuvent également être utilisés pour imputer les questionnaires manquants. Ce modèle de régression utilise les données des patients ayant complétés le questionnaire au temps T pour prédire les valeurs des patients n'ayant pas complétés le questionnaire. Ce modèle utilise la donnée de QdV au score précédent ainsi que les co-variables caractéristiques des patients recueillies à l'inclusion. Ainsi, cette méthode d'imputation peut tenir compte d'un profil de données manquantes MAR.

- Chaîne de Markov

Les méthodes précédentes sont des méthodes déterministes, assignant le même niveau de QdV à deux individus présentant le même profil de QdV est de données manquantes. Une méthode probabiliste utilisant une chaîne de Markov peut également être utilisée pour l'imputation des questionnaires manquants. Cette méthode affecte à un individu dans un état de santé particulier à un temps T la probabilité d'être dans le même état de santé ou dans tout autre état de santé possible au temps suivant. Ces probabilités appelées « probabilité de transition » sont ensuite utilisées pour imputer les valeurs manquantes avec la génération d'un nombre aléatoire entre 0 et 100. Cette méthode est beaucoup plus complexe que les méthodes précédemment énoncées mais elle a l'avantage, contrairement aux méthodes déterministes, de préserver la variabilité des données (Fayers & Machin, 2007).

- Algorithme EM

L'algorithme Espérance-Maximisation (EM) peut être utilisé conjointement avec la méthode de régression ou des chaînes de Markov énoncées précédemment. Il s'agit d'un processus

itératif (comme énoncé pour l'imputation des items selon un modèle IRT). La méthode retenue pour l'imputation est appliquée une première fois. Les données manquantes sont donc remplacées par les valeurs prédites par le modèle. Les paramètres du modèle sont ensuite recalculés avec le nouveau jeu de données. Les données manquantes sont alors remplacées par les nouveaux paramètres du modèle révisé. Ses étapes sont alors répétées jusqu'à ce que les valeurs estimées ne diffèrent pas entre deux estimations successives.

- Imputation « Hot deck »

La méthode d'imputation « Hot deck » sélectionne au hasard, pour chaque individu présentant un score manquant, un score parmi les données observées et remplace le score manquant par cette valeur sectionnée. La sélection du score d'un patient faite au hasard peut être restreinte aux patients présentant les mêmes caractéristiques que celui présentant une donnée manquante.

- Imputations multiples

Comme pour l'imputation des items, des imputations multiples peuvent être utilisées pour les questionnaires totalement manquants. Cette méthode est par ailleurs davantage utilisée pour l'imputation des questionnaires manquants que pour celle des items manquants. Selon le type de données manquantes (intermittentes ou monotones) et le types de variables (continues ou polytomiques), différentes méthodes d'imputation peuvent être utilisées (Tableau 6).

Tableau 6 : Différentes méthodes d'imputations multiples réalisées sous le logiciel SAS

Type de données manquantes	**Type des variables**	**Méthode recommandée**
Monotone	Continues	Régression linéaire
		Predicted Mean Matching
		Score de propension
	Polytomiques ordinales	Régression logistique
Arbitraire	Continues	Monte Carlo par Chaîne de Markov

Comme pour l'imputation multiple des items, l'analyse est réalisée par imputation et les résultats sont ensuite combinés avec l'utilisation de la procédure MIANALYZE du logiciel SAS. L'avantage de cette méthode d'imputation est qu'elle permet de tenir compte de l'incertitude de la valeur de la donnée manquante et permet ainsi de ne pas sous-estimée la variance des données.

Le Tableau 7 suivant résume les caractéristiques des principales méthodes de gestion des données manquantes.

Tableau 7 : Résumé des caractéristiques des principales méthodes de gestion des items ou questionnaires manquants

	Type de données manquantes traité			Tient compte du niveau de :	
	MCAR	MAR	MNAR	QdV du patient	Difficulté de l'item
Données complètes	X				
Données disponibles	X				
Imputation des items					
Personal Mean Score	X			X	
Item Mean Score	X				X
Régression		X		X	X
IRT	X			X	X
Imputations multiples		X		X	X
Imputation des questionnaires					
LOCF	X				
Valeurs arbitraires			X		
Régression		X			X
Markov	X				
Hot deck	X			X	X
Imputations multiples		X		X	

- **Analyse de sensibilité**

Peu de méthodes d'imputations permettent de tenir compte de l'occurrence de données manquantes de type MNAR. Dans le cas de données manquantes de type MNAR, l'idée serait de réaliser, après avoir effectué l'analyse longitudinale selon la méthode statistique retenue, une analyse de sensibilité sur les données manquantes. Ces analyses supplémentaires permettent de vérifier si en modifiant très légèrement un paramètre, les nouveaux résultats diffèrent ou non de ceux obtenus avec l'analyse principale.

A titre d'exemple, pour une analyse longitudinale réalisée selon la méthode du temps jusqu'à détérioration d'un score de QdV, une analyse de sensibilité peut être réalisée en considérant le patients sortie d'étude précocement en détérioration (Bonnetain *et al*, 2010).

La prise en compte des données manquantes non aléatoires peut également être réalisée avec des modèles de type « pattern mixture » (Little & Wang, 1996; Pauler *et al*, 2003; Post *et al*, 2010) en analyse de sensibilité d'un modèle linéaire à effets mixtes basé sur le score. Ce modèle nécessite de construire des « patterns » de données manquantes (Thijs *et al*, 2002). Il permet à la fois de tenir compte de données manquantes intermittentes et de drop-out. Les

analyses et les estimations des paramètres sont alors réalisées par « pattern ». Ces modèles sont cependant peu exploités en QdV en raison de leur complexité. De plus, pour J évaluations, le nombre de pattern possible est de 2^J. Ainsi, si un nombre insuffisants de patients sont présents dans chaque pattern, cela peut poser des difficultés de convergence.

J « dummy » variables (variables indicatrices) sont alors crées égales à 1 si le patient appartient au pattern J, et 0 sinon. Par exemple, pour deux pattern de données manquantes de variables indicatrices respectivement $Pattern_1$ et $Pattern_2$, le modèle contenant des effets fixes bras de traitement, temps et temps interaction traitement peut s'écrire:

$$Y_i^{(t)} = a_1 * Pattern_1 * bras_i + a_2 * Pattern_2 * bras_i + b_1 * Pattern_1 * t + b_2 * Pattern_2 * t + c_1 * Pattern_1 * bras_i * t + c_2 * Pattern_2 * bras_i * t + \varepsilon_i^{(t)}$$

Avec:

- $Y_i^{(t)}$ le score du patient i au temps t
- $bras_i$ est le bras de traitement du patient i (égal à 0 ou 1) ;
- a_j l'effet fixe bras de traitement pour le pattern j et j=1,2 ;
- b_j l'effet fixe temps pour le pattern j ;
- c_j l'effet fixe d'interaction entre le temps et le bras de traitement pour le pattern j ;
- $\varepsilon_i^{(t)}$ le terme d'erreur résiduelle du patient i au temps t suivant une loi normale de moyenne 0 et d'écart-type σ.

Des effets aléatoires peuvent également être ajoutés au modèle.

Les paramètres du modèle sont alors estimés pour chacun de ces « pattern ». Enfin, le résultat final est obtenu en pondérant les trajectoires au cours du temps des différents patterns par leur proportion (Post *et al*, 2010).

6.4. Effet Response Shift

6.4.1. Définition

Un des objectifs majeurs quand on évalue la QdV au cours du temps est de déterminer dans quelle mesure les toxicités dues aux traitements ou la progression de la maladie peuvent affecter le niveau de QdV du patient au cours du temps. Or, la QdV est une mesure subjective puisqu'elle dépend des références internes du patient et de sa propre définition de la QdV. La QdV est aussi un concept dynamique. En effet, on ne peut pas s'assurer que le patient évalue selon les mêmes critères son niveau de QdV au cours du temps. Les attentes et espérances de santé du patient peuvent évoluer au cours du temps du fait du diagnostic de la maladie et de l'adaptation vis-à-vis de la maladie et à la toxicité des traitements. Ce processus est reflété par un effet « Response Shift » ou « changement de réponse » (Howard *et al*, 1979a; Howard *et al*, 1979b).

La Response Shift a été définie par Sprangers et Schwartz en 1999 comme un ensemble de trois composantes (Sprangers & Schwartz, 1999) :

1. un changement dans les références internes du patient (« recalibration »),
2. un changement dans l'importance relative des différentes dimensions de QdV (« reprioritization ») et
3. un changement dans la définition/conceptualisation de la QdV (« reconceptualization »).

La « recalibration » représente un changement quantitatif au niveau de l'échelle de mesure. Supposons que l'on demande à un individu d'évaluer son niveau de fatigue à un instant donné sur une échelle de 0 à 100 ou 100 représente le niveau de fatigue le plus élevé possible. Celui-ci donne un score de 70 sur 100. Cet individu est par la suite traité pour un cancer. On lui demande à nouveau, après avoir été traité par chirurgie puis par chimiothérapie, d'évaluer son niveau de fatigue. Il se peut alors que le patient donne à nouveau un score de 70 sur 100. Cependant, ce nouveau score ne signifie pas que le traitement n'a pas eu d'impact sur son niveau de fatigue. En effet, si on redemande à ce patient, au moment de l'évaluation après traitement, de réévaluer son niveau de fatigue initial, alors ce patient évalue son niveau de fatigue antérieur à 30 sur 100. Le patient a donc revu son jugement initial à la baisse et à « recalibré » son échelle de mesure de la fatigue. Il s'agit bien ici d'un exemple de changement dans les références internes du patient.

Le changement de valeurs ou d'importance relative des différentes dimensions de la QdV (« reprioritization ») correspond à un changement de l'ordre d'importance des différents domaines contribuant à la QdV d'un individu. Prenons l'exemple d'un individu très attaché à son apparence physique. Ainsi, l'activité physique est prioritaire pour cet individu est prend une place importante au regard de sa QdV. Inversement, cet individu porte peu d'importance à sa vie sociale, telle que ses liens avec sa famille et/ou ses amis. On diagnostique chez cet individu un cancer. Après un traitement lourd par chirurgie et par chimiothérapie, cet individu est fortement diminué physiquement. Il se rapproche alors de sa famille et de ses amis et se rend compte que les liens sociaux sont très importants pour lui. Inversement, les activités physiques deviennent moins importantes à ses yeux. Il s'agit ici d'un changement de valeurs des dimensions physiques et sociales de la QdV. Ces changements de valeurs auront des répercussions sur son évaluation de la QdV.

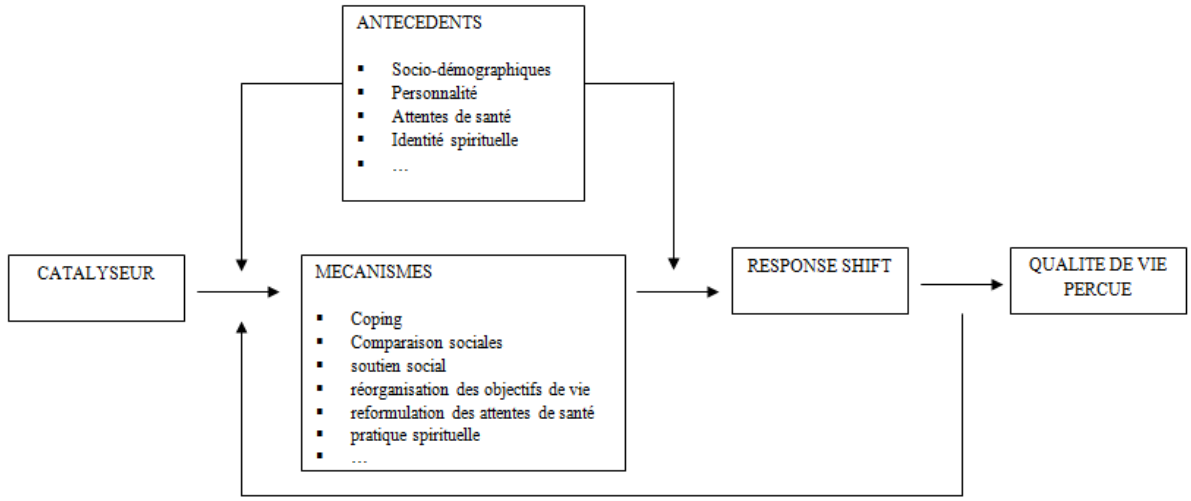
La composante redéfinition ou reconceptuatlisation de la QdV est beaucoup plus conceptuelle, abstraite et difficile à appréhender. A titre d'illustration, un patient apprenant qu'il a un cancer s'engage dans une pratique spirituelle et vient alors en aide à d'autres personnes en situation plus précaire que la sienne. Cette personne entre alors dans un processus de comparaisons sociales. En s'engageant dans une telle pratique spirituelle et en venant en aide à des personnes dont les conditions de vie sont plus précaires et difficiles que les siennes, ce patient va se créer une nouvelle raison de vivre ce qui modifiera sa conceptualisation de la QdV (Sprangers & Schwartz, 1999).

La mesure dans laquelle les trois composantes de la Response Shift sont clairement distinctes ou bien au contraire interconnectées est toujours en question (Sprangers & Schwartz, 1999). Cependant, il semble cohérent que l'occurrence d'une recalibration pour une échelle donnée pourrait refléter une modification en amont de l'importance des dimensions liées à cette échelle. Par exemple, une personne fatiguée recalibrant son échelle de fatigue à la baisse pourrait également revoir l'importance de sa condition physique à la baisse.

Schwartz *et al.* ont ainsi proposé un modèle théorique permettant d'illustrer le processus d'occurrence de l'effet Response Shift et son impact sur l'évaluation longitudinale de la QdV (Sprangers & Schwartz, 1999). Ainsi, les auteurs proposent un modèle à cinq composantes (Figure 7) :

- Un « catalyseur » correspondant au changement de l'état de santé du patient. Par exemple, le diagnostic d'un cancer.

- Les « antécédents » du patient, incluant ses caractéristiques socio-démographiques, sa personnalité et ses caractéristiques spirituelles.
- Les « mécanismes », faisant référence au processus comportemental et cognitif dont le patient fait preuve pour s'accommoder au catalyseur. Ces mécanismes peuvent inclure des stratégies de coping, de réorganisation des objectifs de vie et de reformulation des attentes et espérances de santé. Ces mécanismes sont créés suite au catalyseur et sont influencés par les antécédents du patient.
- La résultante serait l'occurrence de l'effet Reponse Shift.
- Et enfin, ceci influencerait la QdV perçue par le patient.



<u>Figure 7</u> : Modèle théorique de la Response Shift et son impact sur la QdV perçue (Sprangers & Schwartz, 1999)

L'occurrence d'un effet Response Shift nécessite donc, en amont, un changement de l'état de santé du patient, correspondant par exemple au diagnostic d'un cancer. Néanmoins, un effet Response Shift négligeable peut être mis en évidence même en l'absence de ce catalyseur (Ahmed *et al*, 2014).

6.4.2. <u>Impact de la Response Shift</u>

L'occurrence d'un effet Response Shift peut avoir des conséquences importantes sur l'analyse longitudinale de la QdV (Ring *et al*, 2005; Wilson, 1999). En effet, l'occurrence d'un tel effet peut masquer, sous ou sur estimer l'effet dû au traitement. Par exemple, si on demande à un

patient d'évaluer son niveau de fatigue à deux occasions : au moment du diagnostic d'un cancer et un mois après avoir été opéré. La différence entre les deux mesures peut être nulle ou non cliniquement pertinente. Cependant, si on demande à ce même patient lors de la mesure effectuée à un mois post-chirurgie, de réévaluer également son niveau de fatigue initial au moment du diagnostic, il se peut que celui-ci revoie son jugement à la baisse et rapporte un niveau de fatigue plus faible que celui donné préalablement. Dans ce cas, la comparaison entre son niveau de fatigue après chirurgie et celui donné de façon rétrospective lors de la même mesure de son niveau de fatigue à l'inclusion peut être significative et indiquer une augmentation de son niveau de fatigue après chirurgie. Or, si on ne tient pas compte de la mesure rétrospective, on conclura, a priori à tort, que la chirurgie n'a pas eu d'effet sur le niveau de fatigue du patient. C'est dans ce sens que l'effet Response Shift est souvent considéré comme un biais de mesure.

Dans le cadre d'un essai clinique, une occurrence différentielle d'un effet Response Shift entre les deux bras de traitement rendra les résultats inexploitables si cet effet n'est pas pris en compte de façon adéquate.

Un effet Response Shift a ainsi été démontré à de multiples occasions et dans diverses situations. En cancérologie, cet effet a par ailleurs été mis en évidence dans le cancer du sein en situation adjuvante (Andrykowski *et al*, 2009; Dabakuyo *et al*, 2013) ou dans le cancer de la prostate en situation avancée (Rees *et al*, 2005).

Ce phénomène est désormais bien reconnu dans le domaine de la QdV. Il est donc nécessaire de pouvoir mettre en évidence un tel effet et en tenir compte de façon adéquate dans l'analyse longitudinale.

6.4.3. Méthodes pour caractériser l'occurrence d'un effet Response Shift

Différentes méthodes ont été proposées pour mettre en évidence l'occurrence d'un effet Response Shift (Ahmed *et al*, 2005; Boucekine *et al*, 2013; Li & Rapkin, 2009; Lix *et al*, 2013; Oort, 2005; Schwartz & Sprangers, 1999). Toutes ces méthodes ont leurs avantages et leurs inconvénients et permettent de mettre en évidence une, deux ou les trois composantes de la Response Shift.

Deux stratégies sont possibles pour détecter l'occurrence d'un effet Response Shift :

- soit a priori, en prévoyant de mesurer la Response Shift dans le design de l'étude avec la passation de questionnaires supplémentaires,
- soit a posteriori, en développant des méthodes statistiques pour mettre en évidence l'occurrence d'un tel effet sur les données de QdV recueillies.

- **Méthode « Then-test »**

La méthode « Then-test » est une des méthodes les plus anciennes pour mettre en évidence l'occurrence de la Response Shift et est encore souvent utilisée à ce jour (Schwartz & Sprangers, 1999; Sprangers *et al*, 1999b). Elle consiste à introduire dans le design de l'étude une mesure rétrospective du niveau de QdV antérieur du patient. Le test consiste à demander aux patients après traitement, par exemple, d'évaluer leur niveau de QdV actuel (mesure « post-test ») mais aussi leur niveau de QdV avant traitement (au moment du pré-test) de façon rétrospective. Cette méthode est basée sur l'hypothèse que les évaluations post-test et then-test sont réalisées selon les mêmes références internes pour le patient puisque elles sont effectuées au même temps d'évaluation. La composante recalibration de la Response Shift serait ainsi mise en évidence en comparant la différence de moyennes entre le score obtenu lors du pré-test et celui obtenu lors du then-test. La différence entre la mesure then-test et la mesure post-test permettrait d'évaluer l'effet traitement, tout en tenant compte de l'occurrence de la recalibration, i.e. le vrai changement.

Cette méthode est simple à implémenter puisqu'elle nécessite uniquement la passation d'un questionnaire supplémentaire. Cependant, certains inconvénients ont été mis en évidence (McPhail & Haines, 2010). En effet, la mesure rétrospective nécessite que le patient se souvienne de son niveau de QdV antérieur. Si la mesure rétrospective est relativement éloignée de la mesure initiale, il peut être difficile pour le patient de se souvenir très clairement de son état de santé initial. Ainsi, un biais de mémoire peut avoir lieu. D'autre part, le patient peut enjoliver sa perception de son niveau de QdV antérieur dans un souci de désirabilité sociale par son médecin.

En 2010, Schwartz *et al.* ont proposées des recommandations pour rendre plus rigoureuses les recherches sur la Response Shift utilisant la méthode « Then-test » (Schwartz & Sprangers, 2010). Ces recommandations concernent le design de l'étude, la construction du then-test, ainsi que des recommandations au niveau des méthodes statistiques à appliquer et au niveau

de l'interprétation des résultats. Ces recommandations inclues la nécessité d'introduire un groupe contrôle dans l'étude qui ne serait pas susceptible de présenter un effet Response Shift. Par exemple, les patients dans un état de santé stable ne devraient pas présenter d'effet Response Shift puisque celle-ci est catalyseur dépendant, i.e. dépend d'un changement de l'état de santé de l'individu. Le questionnaire de QdV administré aux patients doit également être correctement adapté pour la mesure rétrospective. La présentation du questionnaire doit être modifiée afin de demander clairement au patient de se souvenir de son niveau de QdV antérieur et non pas d'évaluer son niveau de QdV actuel. Concernant l'analyse statistique, il est essentiel de reporter la différence de moyenne entre la mesure then-test et la mesure pré-test et de quantifier la taille (la magnitude) de l'effet Response shift par l' « Effect Size ». Cela permettrait en outre de pouvoir comparer les résultats entre les essais et de réaliser une méta-analyse. De plus, il est recommandé de tester l'invariance de la mesure then-test, i.e. de vérifier l'hypothèse que le patient évalue de la même façon sa QdV lors des deux mesures then-test et post-test (Nolte *et al*, 2009).

- **Méthode « Ideal Scale Approach »**

L'approche de l'échelle idéale (« Ideal Scale Apporach ») consiste à demander au patient d'évaluer sur une même échelle son niveau de QdV actuel ainsi que son niveau de QdV idéal (Schwartz & Sprangers, 1999). En appliquant cette double évaluation à chaque temps de mesure, un changement de références internes pourra être détecté par un changement de score idéal.

- **Méthode « Successive Comparison Approach »**

La méthode des comparaisons successives (« Successive Comparison Approach ») consiste à demander au patient à plusieurs reprises au cours de l'étude de classer par ordre d'importance certaines dimensions ou domaines de QdV (Schwartz & Sprangers, 1999). Ainsi, un changement de valeurs sera mis en évidence par un changement de l'ordre d'importance des différentes dimensions ou domaines.

- **Schedule for the Evaluation of Individual Quality of Life**

Le questionnaire SEIQoL (« Schedule for the Evaluation of Individual Quality of Life ») est un questionnaire souvent utilisé pour évaluer de façon individuelle la QdV des patients (Hickey *et al*, 1996; O'Boyle *et al*, 1992). Il doit être rempli lors d'un entretien avec un psychologue ou un attaché de recherche clinique. On demande ainsi au patient de sélectionner les cinq domaines les plus importants pour sa QdV puis d'évaluer son niveau actuel dans chacun de ces domaines sur une échelle visuelle analogique. Un indice global est généré résumant la satisfaction globale de chaque domaine et tenant compte de leur importance relative. La composante « changements de valeurs » de la Response Shift sera alors reflété par un changement de l'importance relative des différentes dimensions de QdV. Une des limites majeures de ce questionnaire est qu'il nécessite un entretien avec un attaché de recherche clinique ou un psychologue pour pouvoir être complété.

Différentes méthodes statistiques ont également été proposées pour mettre en évidence l'occurrence d'une ou plusieurs composantes de la Response Shift (Ahmed *et al*, 2005; Barclay-Goddard *et al*, 2009; Lix *et al*, 2013; Oort, 2005).

- **Analyses factorielles**

Des analyses de covariances ont été proposées pour mettre en évidence l'occurrence de la Response Shift en utilisant une analyse factorielle confirmatoire (Ahmed *et al*, 2005; Schmitt, 1982). Ces analyses permettraient d'identifier les composantes reconceptualisation et « changements de valeurs » par un changement de structure factorielle et des « factor loadings » au cours du temps.

- **Modèles à équations structurelles**

Une méthode alternative pour mettre en évidence l'occurrence de la Response Shift avec des analyses factorielles exploratoires a été proposée par Oort en 2005 (Oort, 2005). L'auteur a proposé une procédure séquentielle pour mettre en évidence l'occurrence de chaque composante de la Response Shift à partir des modèles à équations structurelles (« Structural Equation Modeling », SEM) (Oort, 2005).

Supposons que le niveau de QdV du patient est évalué par des variables observées multiples. Le modèle pour les scores observés d'un individu i donné peut s'écrire :

$$Y_i = \tau + \Gamma\xi_i + \zeta_i$$

où Y_i est le vecteur des scores observés, ξ_i correspond au vecteur des facteurs communs non observés et ζ_i est un vecteur des facteurs résiduels non observés. La matrice Γ contient les « factors loadings » et le vecteur τ les intercepts.

Les trois composantes de la Response Shift peuvent ainsi être détectées selon ce modèle et la procédure de Oort. L'interprétation des différentes composantes est la suivante :

- La **recalibration non uniforme** concerne une différence au niveau de la variance des erreurs entre deux temps de mesures. Les facteurs résiduels représentent toute la variance dans les scores qui n'est pas expliquée par les facteurs communs.

- La **recalibration uniforme** correspond à un changement d'intercept. L'intercept représente la « facilité » d'avoir un score élevé pour l'échelle considérée. Pour illustration, considérons deux scores, un de fatigue et le second de douleur, construit sur une même échelle standardisée de 0 à 100 tel qu'un score élevé représente un niveau de fatigue/douleur élevé. Si l'intercept de la fatigue est plus faible que celui de la douleur, alors il est plus facile de présenter un score élevé pour la douleur que pour la fatigue. En d'autres termes, étant donné un certain niveau du trait latent (par exemple, état symptomatique), les patients présenteront davantage de douleur que de fatigue.

- Le **changement de valeurs** et la **reconceptualisation** concernent les factor loadings. Les factors loadings montrent la contribution de chaque variable pour le facteur commun (la variable latente). Par exemple, si le factor loading de la douleur est plus élevé que celui de la fatigue, alors la douleur contribue davantage au niveau symptomatique du patient (i.e. devient plus importante) que le niveau de fatigue.

Dans la terminologie des modèles IRT :

- le « facteur loading » peut être interprété comme le paramètre de discrimination de l'item
- et l'intercept comme le paramètre de difficulté de l'item.

La première étape de la procédure de Oort consiste à concevoir un modèle de mesure qui aurait une interprétation claire et qui aurait une bonne adéquation aux moyennes et

covariances observées. Dans cette étape, tout type de Reponse Shift est autorisé mais le vrai changement est fixé à 0 (Modèle 1).

La seconde étape consiste à réaliser un test global de l'occurrence d'un effet Response Shift. Un modèle est alors construit, basé sur ce modèle de mesure, mais dans lequel aucun effet Response Shift n'est toléré (Modèle 2). Le vrai changement est alors estimé. On fixe donc dans ce modèle un certain nombre de contraintes. Les Modèles 1 et 2 sont alors comparés et si la différence entre les deux modèles est significative, on conclut qu'un effet Response Shift a lieu.

Si l'étape 2 a conclu à l'occurrence d'un effet Response Shift, on passe à l'étape 3. A partir du Modèle 2, on relâche alors une à une les contraintes afin de permettre l'occurrence d'un effet Reponse Shift (Modèle 3). Nolte *et al*. (Nolte *et al*, 2009) ont proposé d'introduire de manière séquentielle et dans un ordre défini chaque type de Response Shift: 1) d'abord on teste la présence d'une recalibration non uniforme ; 2) puis la recalibration uniforme ; 3) le changement de valeurs ou d'importance relative des différentes dimensions de QdV ; 4) et enfin la reconceptualisation de la QdV.
La procédure s'arrête lorsque l'information apportée par le nouveau modèle n'est plus significative par rapport à celle du Modèle 2, on obtient alors un modèle final (Modèle 4), dans lequel l'effet Reponse Shift et le vrai changement sont estimés.

Cette méthode à l'avantage de pouvoir mettre en évidence chaque composante de la Response Shift. Elle a été utilisée à de nombreuses reprises (King-Kallimanis *et al*, 2011; King-Kallimanis *et al*, 2009; Oort *et al*, 2005). Cependant, jusqu'à ce jour et à ma connaissance, cette procédure n'est appliquée qu'au questionnaire générique SF-36 respectant une structure particulière en deux macro domaines (physique et mental).

La procédure de Oort a également été reprise par Notle *et al.* (Nolte *et al*, 2009) pour tester l'invariance de la mesure then-test. Cette méthode est alors appliquée entre les mesures then-test et post-test. L'objectif étant de tester la capacité de la Méthode Then-test à bien mettre en évidence un effet Response shift, et non un autre biais de mesure comme un biais de mémoire. Si le then-test est bien capable de capter cet effet Response Shift seul, la comparaison entre la mesure then-test et la mesure post-test selon la méthode de Oort ne mettra donc pas en évidence de changements aux niveaux des paramètres d'intercept, des « factor loadings » ou des variances résiduelles, mais uniquement un possible effet « vrai changement ».

- **Differential Item Functioning via les modèles IRT**

Les DIF correspondent à un fonctionnement différentiel de l'item entre différents groupes d'individus. Ainsi, si un DIF est démontré entre deux temps de mesure pour une même population d'individus, cela sous-entend l'occurrence d'un effet Response Shift. Les modèles IRT semblent avoir de bons potentiels pour pouvoir démontrer un effet Response Shift (Barclay-Goddard *et al*, 2009).

Cependant, à ce jour, les modèles IRT n'ont encore pas été explorés à proprement parlé pour pouvoir caractériser l'occurrence de l'effet Response Shift. Des recherches semblent donc nécessaires pour pouvoir déterminer la capacité de ces modèles à mettre en évidence l'occurrence d'un tel effet.

- **Analyse discriminante et modèle logistique**

Lix *et al.* ont proposé en 2012 deux méthodes pour mettre en évidence la composante « changement de valeurs » de la Response Shift (Lix *et al*, 2013) utilisant l'analyse discriminante ainsi que les modèles logistiques. Ces méthodes sont basées sur l'hypothèse que le changement de valeurs résulte d'un changement statistiquement significatif entre deux groupes de patients tel qu'un groupe traitement versus un groupe contrôle, des patients avec beaucoup de symptômes versus peu, des patients avec une maladie active versus en rémission. L'analyse discriminante teste un changement de poids des différentes associations entre les domaines de QdV tout en tenant compte de la variable de groupe. La seconde méthode teste un changement de rangs des différents domaines. Cependant, l'exploration de ces méthodes n'a pas montré de résultats consistants entre les deux approches. D'autre part, elle requiert la constitution de deux groupes de patients : le premier groupe susceptible de présenter un effet changement de valeurs de la Response Shift, et le second groupe dans un état de santé stable.

A ce jour, il n'existe pas de recommandations sur les méthodes statistiques à investiguer pour mettre en évidence l'occurrence de la Response Shift. Ainsi, il parait indispensable d'investiguer différentes approches statistiques sur un même échantillon afin de comparer les résultats en terme de profils de Response Shift. Des résultats similaires obtenus avec différentes approches permettraient ainsi d'évaluer la validité convergente entre ces approches sur un même échantillon (Barclay-Goddard *et al*, 2009). D'autre part, bien que l'occurrence

potentielle d'un effet Response Shift soit désormais bien établie, peu d'investigation ont été faites d'un point de vue longitudinal (avec plus de deux temps de mesure) et aucune méthode statistique pour l'analyse longitudinale de la QdV ne permet de prendre en compte l'occurrence d'un tel effet à ce jour.

III. OBJECTIFS

Dans les essais cliniques en cancérologie, la QdV est considérée comme le second critère de jugement principal par l'ASCO et la FDA en l'absence d'effet sur la survie globale (Beitz *et al*, 1996). De plus, la discussion actuelle porte de plus en plus sur la prise en compte de la QdV comme co-critère de jugement principal avec un critère tumoral comme la survie sans progression, notamment en situation avancée ou métastatique (Bonnetain *et al*, 2012). Cependant, les résultats des analyses de QdV restent encore peu pris en compte par les cliniciens pour changer les standards thérapeutiques de prise en charge des patients. Ce manque de considération de la QdV est principalement dû à la complexité de la mesure et de l'analyse longitudinale des données de QdV. En effet, il n'existe pas de consensus autour de la définition de la QdV. De plus, la QdV est un concept subjectif et dynamique et peut être impacté par l'occurrence de données manquantes potentiellement informatives de l'état de santé du patient. De nombreuses méthodes d'analyse longitudinale existent mais à ce jour aucun standard n'a pu être établi. L'analyse longitudinale doit tenir compte de l'occurrence de données manquantes intermittentes et monotones et d'un possible effet Response Shift. D'autre part, les méthodes statistiques proposées sont souvent sophistiquées et les résultats peu accessibles aux cliniciens. Il parait donc indispensable de proposer une approche statistique pertinente pour l'analyse longitudinale, proposant des résultats facilement compréhensible pour les cliniciens et pouvant prendre en compte l'occurrence d'un possible effet Response Shift ainsi que les données manquantes.

Des méthodes d'analyses statistiques plus ou moins sophistiquées ont été proposées pour l'analyse longitudinale des données de QdV (Douglas, 1999; Fairclough, 2010; Fayers & Machin, 2007). Depuis quelques années, la méthode du temps jusqu'à détérioration d'un score de QdV est largement utilisée comme modalité d'analyse longitudinale de la QdV dans les essais cliniques de phase III en cancérologie (Bonnetain *et al*, 2010; Burris *et al*, 2013; Gourgou-Bourgade *et al*, 2013; Kabbinavar *et al*, 2008). Cette méthode dite de « temps jusqu'à évènement » est une méthode de type analyse de survie qui requiert donc une définition de son évènement, soit la détérioration. Cette approche est assez attractive pour les cliniciens puisque les résultats sont facilement compréhensibles et peuvent être résumés en termes d'Hazard Ratio et de médiane de survie par bras de traitement. Différentes définitions ont déjà été proposées dépendant du score de référence, de la différence minimale

cliniquement importante, de la prise en compte ou non des données manquantes et intégrant ou non le décès dans la définition de l'évènement (Bonnetain *et al*, 2010; Hamidou *et al*, 2011). Généralement, les données manquantes intermittentes et monotones ayant lieu au cours du suivi ne sont pas prises en compte dans l'analyse du TJD, considérant que le niveau de QdV du patient est resté constant depuis la dernière mesure disponible.

La multiplicité des définitions de TJD possibles implique certaines recommandations pour l'analyse longitudinale des essais cliniques selon les situations thérapeutiques afin de pouvoir comparer les résultats entre les essais. Ces définitions doivent également pouvoir être adaptées en présence d'un effet Response Shift. De plus, compte tenu du nombre croissant d'essais cliniques utilisant cette méthode, il parait nécessaire d'implémenter les différentes définitions proposées sous un logiciel statistique accessible permettant de réaliser facilement ces analyses. L'impact des données manquantes non MCAR sur cette approche doit également être étudié. Enfin, la QdV reste encore peu exploitée dans les essais de phase précoce, en particulier les essais de phase I. Le TJD pourrait être une méthode adaptée à l'analyse de la QdV dans ces essais tout en tenant compte de l'occurrence de toxicités évaluées selon la grille NCI-CTC AE.

D'autres méthodes statistiques pour l'analyse longitudinale de données de QdV existent. Le modèle linéaire à effets mixtes basé sur le score reste à ce jour la méthode statistique la plus souvent utilisée mais elle pourrait ne pas être la méthode la plus adaptée à des données issues des questionnaires de QdV (Fairclough, 2010; Pan *et al*, 2012). Les modèles IRT sont potentiellement bien adaptés aux données issues des questionnaires de QdV. Néanmoins, ces modèles sont complexes et leur extension à l'analyse longitudinale reste peu exploitée (Douglas, 1999). Plusieurs études de simulations ont été menées sur la comparaison de ces deux modèles pour l'analyse longitudinale de la QdV (Blanchin *et al*, 2011a; Blanchin *et al*, 2011b; de Bock *et al*, 2013; de Bock *et al*, 2014). Ces études présentent certaines limites méthodologiques. En effet, elles ont toutes été menées sur des échelles construites avec des items dichotomiques. Le questionnaire EORTC QLQ-C30, comme la majorité des questionnaires utilisés en cancérologie, contient des items polytomiques. Ces études se sont également restreintes à trois temps de mesure alors que davantage d'évaluations sont généralement planifiées dans les essais. De plus, ces études se sont également focalisées sur un effet temps ou un effet bras de traitement. Or, dans les essais cliniques randomisés, le niveau de QdV est supposé être égal dans les deux bras de traitement à l'inclusion. Pour détecter un effet de traitement différent entre les deux bras de traitement au cours du temps, il

parait donc pertinent de s'intéresser à un effet d'interaction entre le temps et le bras de traitement. Ainsi, il parait indispensable de comparer les trois approches (TJD, modèle linéaire à effets mixtes basé sur le score, IRT) sur des données simulées en se focalisant sur la structure du questionnaire QLQ-C30 de l'EORTC et selon des scénarios proches des conditions réelles des essais cliniques. Cette comparaison permettrait de déterminer quelles sont les méthodes pertinentes et robustes pour l'analyse longitudinale de la QdV selon les situations thérapeutiques et le design de l'étude.

Enfin, un dernier enjeu important de l'analyse longitudinale est l'occurrence potentielle de l'effet Response Shift. L'occurrence potentielle d'un tel phénomène est clairement reconnue et établie (Sprangers & Schwartz, 1999). De nombreuses méthodes ont été proposées pour caractériser l'occurrence d'un tel effet (Li & Rapkin, 2009; Lix *et al*, 2013; Oort, 2005; Schwartz & Sprangers, 1999). Ces méthodes ont toutes leurs avantages et leurs inconvénients. Une des méthodes les plus utilisées est la méthode then-test qui consiste à introduire, dans le design de l'étude une mesure rétrospective de la QdV (Schwartz & Sprangers, 1999). Cependant, cette méthode peut être affectée par un biais de mémoire et de désirabilité sociale (McPhail & Haines, 2010). Elle nécessite également la passation de questionnaires supplémentaires et doit être prévue au moment du design de l'étude. Les modèles à équations structurelles à travers la procédure de Oort ont également été proposées comme méthode statistique pour mettre en évidence les trois composantes de la Response Shift (Oort, 2005). Cette méthode a été appliquée à de multiples occasions, mais uniquement sur le questionnaire générique SF-36 (King-Kallimanis *et al*, 2011; King-Kallimanis *et al*, 2009; Oort *et al*, 2005). Il parait donc indispensable d'explorer la capacité de cette méthode à mettre en évidence un effet Response Shift sur les questionnaires spécifiques du cancer. L'investigation de différentes approches statistiques parait ainsi nécessaire pour tester la cohérence entre les résultats. De plus, à ce jour, peu d'études ont investigué l'occurrence de l'effet Response Shift de façon longitudinale, i.e. avec plus de deux temps de mesure. Cet effet étant susceptible de varier au cours du temps, une analyse longitudinale est par ailleurs requise.

Dans cette optique, les objectifs de ce travail ont été :

1. D'investiguer la méthode du temps jusqu'à détérioration d'un score de QdV en tant que modalité d'analyse longitudinale de la QdV et ceci en :
    i. Proposant un standard pour l'analyse longitudinale selon la méthode du TJD selon les situations thérapeutiques et l'occurrence ou non d'un effet Response Shit ;
    ii. Implémentant l'ensemble des définitions possibles du TJD sous un logiciel de statistiques dans le but de permettre une large utilisation de la méthode et une standardisation des définitions ;
    iii. Proposant une méthode statistique pour tenir compte de l'occurrence des données manquantes au cours du suivi conjointement avec la méthode du TJD.
2. D'étudier la méthode du TJD dans un essai de phase I et d'explorer l'ajout de la QdV dans la détermination de la dose recommandée pour les essais de phase II.
3. De comparer trois méthodes statistiques pour l'analyse longitudinale de la QdV par le biais de simulations avec plusieurs scénarios et d'étudier l'impact des données manquantes sur ces différentes stratégies.
4. De caractériser l'occurrence de l'effet Response Shift dans une étude longitudinale de QdV de patients ayant un cancer du sein par le biais de deux approches statistiques :
    i. Conjointement avec la méthode « then-test »
    ii. En se basant uniquement sur les mesures prospectives.

IV. TRAVAUX REALISES

1. Challenges de l'analyse statistique des données de QdV dans les essais cliniques en cancérologie

Article en révision dans *Journal of Clinical Oncology*

Statistical Challenges in the Analysis of Health-Related Quality of Life in Cancer Clinical Trials

Franck Bonnetain (1, 2 ,3 , 5), Frédéric Fiteni (1,2,4), Amélie Anota (1,2,3)

1 University Hospital of Besançon, Methodology and Quality of Life in Oncology Unit, Besançon, France

2 EA 3181 University of Franche-Comté, Besançon, France

3 The French National Platform Quality of Life and Cancer, Besançon France

4 University Hospital of Besançon, Department of Medical Oncology, Besançon, France

5 EORTC QOL Group, Brussels, Belgium

Key words: Health related Quality of life, longitudinal analysis, Methodology, endpoint

Corresponding Author:

Professor Franck Bonnetain

Head of methodological and quality of life unit in oncology (EA3181) & associated head of quality of life and cancer clinical research platform CHU Besançon,

2 place Saint Jacques 25 030 Besançon, France, Tel: + 33 (0)3 81 21 92 06,

E-mail:
franck.bonnetain@univ-fcomte.fr; fbonnetain@chu-besancon.fr

The authors thank Dr Magdalena Benetkiewicz and Molly Votaw for their medical editorial assistance in manuscript preparation and review.

Introduction

Endpoints refer to clinical and biological measurements that assess efficacy of therapeutic strategies. As the American Society of Clinical Oncology states, active treatment in cancer is generally undertaken with the goals of providing improved quantity and/or quality of patient survival. The Food and Drug Administration (FDA) considers OS benefit as the foundation for the approval of new anticancer drugs in the United States. Nevertheless, with the increasing number of effective salvage treatments available in many types of cancers, trials require a large number of patients and/or long-term follow-up to observe the required number of events with desired level of statistical power.[1] Thus, endpoints that can be assessed earlier, such as progression-free survival, are frequently used as primary endpoints instead of OS. Moreover, the use of PFS may be appropriate in particular disease settings and has been accepted by regulatory authorities on many occasions [2,3]. However, these exhibit important limitations, notably heterogeneity of their definitions across studies and lack of validation as surrogates for OS in all cancer localizations, tempering conclusion about clinical benefit.[1,4]

In this context, Health related Quality of Life (HRQoL) is recognized as a component endpoint for cancer therapy approvals by the American Society of Clinical Oncology and the FDA. In order to make this possible, guidelines for methods of analyzing and reporting longitudinal changes in HRQoL scores in clinically meaningful ways are mandatory.

The goal of this paper is to bring to light some challenges associated with statistical analyses of HRQoL in cancer clinical trials.

HRQoL assessment and questionnaire validation

HRQoL reflects the patient-perceived evaluation of one's health, including physical, emotional, and social dimensions as well as symptoms due to disease or treatment.

In oncology, many self-completion HRQoL questionnaires have been developed and validated.

The European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire QLQ-C30 and the Functional Assessment of Cancer Therapy- General (FACT-G) are the most widely used cancer specific instruments in phase III clinical trials.[5,6] According to cancer sites or treatment modalities, supplementary modules to the QLQ-C30[7] were developed, such as the 23-item BR23 module[8] for breast cancer and the 13-item LC13 module[9] for lung cancer. In the context of the FACT group, specific questionnaires were also developed, for instance, for cancer sites containing the FACT-G items and dimension-specific to the localization such as the FACT-B questionnaire[10] for breast cancer and the FACT-L questionnaire[11] for lung cancer.

Prior to utilization, questionnaires must demonstrate some psychometrics properties such as internal reliability, reproducibility, and responsiveness to changes. Moreover, transcultural adaptation is also required as translation is not sufficient. To be qualified as relevant and reliable tools, these properties are checked through classical and item response theory (IRT) analyses[12].

At this time, the foremost challenge would be to promote, through cancer site and treatment modalities, guidelines for selecting the best questionnaires allowing for direct comparison of results across trials. As an example, two recent phase III clinical trials investigated the impact of the addiction of bevacizumab to the standard therapy of newly diagnosed glioblastomas.[13,14] Both trials used the EORTC QLQ-C30 and its BN20 module in HRQoL analysis. These two similar clinical trials could then be directly compared.

On the contrary, two recent phase III clinical trials, OPTIMAL and LUX-Lung 3, focused on patients with advanced non-small cell lung cancer with EGFR mutations and investigated an

inhibitor of EGFR (erlotinib for OPTIMAL trial and afatinib for LUX-Lung 3 trial respectively) versus chemotherapy [15,16]. HRQoL was assessed by the FACT-L questionnaire in OPTIMAL trial while EORTC QLQ-C30 and LC13 module were used in LUX-Lung 3 trial. These two clinical trials could hardly be compared since EORTC and FACT questionnaires for lung cancer do not contain the same dimensions in regard to impact of lung cancer on HRQoL.

While some guidelines could be proposed for questionnaire selection, it is also still necessary to develop some new tools to evaluate HRQoL. Already validated questionnaires may not be adapted to newly targeted biotherapy agents which can induce some long-term moderate toxicities [17]. In this context, the development of a Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events has been proposed [18,19]. Indeed, along with changing treatment regimens and therapeutic approaches, new toxicities could occur that are not captured by current HRQoL tools. Thus, new HRQoL questionnaires capturing such toxicities should be created.

Moreover, HRQoL is by nature a subjective endpoint integrating psychological dimensions of the patient's feelings about the disease and the impact of treatment. Indeed, a placebo effect could occur which constitutes a limitation of HRQoL assessment. In this way and according to the aim of the trials, it could be more relevant to target primary analyses of HRQoL on the most objective dimensions such as physical and/or symptomatic dimensions.

In oncology, most phase III clinical trials include HRQoL measures through these validated questionnaires in order to ensure the clinical benefit for the patients[20]. However, the analysis of this data still needs some recommendations.

Statistical analysis of longitudinal HRQoL data

In oncology clinical trials, patients' HRQoL levels are evaluated at several time intervals during the trial. Longitudinal analysis of this data is thus required to take into account correlation between HRQoL measures. The planned time points and follow-up of HRQoL measures depend on study design, objectives, and treatment modalities. The number of HRQoL measures and intervals between two consecutive measures may vary from one study to another. HRQoL is often captured until tumor progression, which could constitute a selection of population (potential bias). Indeed, few HRQoL measures would be available for patients with early tumor progression while these patients are more likely to present a low HRQoL level. Thus, it would be more appropriate to measure HRQoL until death.

Such longitudinally collected data has been often underexploited, giving HRQoL mean by treatment arm with the number of patients deteriorated or improved during the treatment.[21] These analyses are only descriptive and refer to a subset of patients with high HRQoL levels[22] since drop out that could reflect deterioration in patients' health status is ignored. Moreover, in order to use all HRQoL information, longitudinal modelling should be the statistical approach to retain.

- Minimal Clinically important difference

Prior to longitudinal HRQoL data analysis, targeted dimensions should be determined a priori in the protocol (to control false negative results) as well as the minimal clinically important difference (MCID).[23] For example, in LUX-LUNG3 trials, pre-specified HRQoL measures of interest were cough and dyspnea dimensions of LC13 module and pain dimension of QLQ-C30 questionnaire [16]. The MCID represents the smallest changes/differences in HRQoL score, which is perceived as clinically important. For the EORTC questionnaires, a 5-point to a 10-point difference in scores could be considered as the MCID.[24,25] Recent studies have shown a

consistent cut off through cancer sites.[26,27] In context of the FACT, a 7-point difference in the global score is generally considered as the MCID[6,28-30]. The MCIDs are approximate and will vary by a few points across cancers and patient populations. There are no guidelines at this time to determine the MCID[31]. An alternative is to consider half of the standard deviation as the MCID in the absence of some recommendations[32].

- linear mixed model for repeated measures

The most widely used method to analyze longitudinal HRQoL data is the linear mixed model for repeated measures (LMM)[33]. This model can estimate time effect (i.e. change of HRQoL score over time), treatment arm effect (i.e. difference between treatment arms at baseline), and an interaction between treatment arm and time, reflecting a different evolution of the two treatment arms over time. The MCID, instead of *P*-value alone, is a key component of the clinically meaningful results. Some random effects are often added to the model in order to reflect inter-patient variability. The structure of the correlation matrix between HRQoL measures must be specified in order to take into account the type of the correlation evolution between HRQoL measures over time. The most commonly used are the unstructured matrix (most general structure), the heterogeneous first-order autoregressive (correlation between HRQoL measure decreases over time) and the heterogeneous compound symmetry (constant correlation between HRQoL measures over time). The Akaike information criterion is used to determine the best structure. In case of non-random missing data, "pattern mixture models" should be performed as sensitivity analysis to deal with missing data patterns.[34] Finally, while these models are robust, heterogeneous underlying construction is challenging, and results are sometimes not meaningful for clinicians. These models generally require a normal distribution of the studied score. Dimensions evaluated with few items (as symptomatic dimensions of the EORTC HRQoL questionnaires) could then not be adapted to LMM models

since the score generated will not be normally distributed. Conversely, each dimension of the FACT HRQoL questionnaires contains at least six items with 5 response categories per item. The FACT scores could thus be better indicated than EORTC scores for using LMM approach. In the randomized open-label phase III AURELIA trial, adding bevacizumab in patients with platinum-resistant ovarian cancer to single-agent chemotherapy significantly improved the primary endpoint of PFS with no significant difference in OS. A LMM model was performed to analyze longitudinal HRQoL data as a secondary endpoint in this trial [35]. More patients with bevacizumab than with chemotherapy alone achieved a ≥15% improvement in abdominal/GI symptoms at week 8/9 (primary HRQoL endpoint, 21.9% *v* 9.3%; difference, 12.7%; 95% CI, 4.4 to 20.9; $P<0.002$). In agreement with published guidelines[36,37] and CONSORT Patient-Reported Outcome[38], this study used LMM statistical analysis with an adequate reporting of the results and a clear a priori definition of the objectives, hypotheses and of the population data set, as well as analysis of missing data, and sensitivity analysis. Moreover, clinically meaningful HRQoL results were convergent with the results of the primary endpoint (PFS).

- Time to HRQoL score deterioration

Survival estimations, such time to HRQoL score deterioration (TTD) approaches, began to be extensively used in oncology phase III clinical trials.[13,39-42]. These models have the advantage of obtaining clinically meaningful results. In adjuvant settings, the most adapted definition for TTD could be the time from inclusion-randomization in the study to a first deterioration of at least one MCID unit, as compared to the baseline score.[43] In advanced or metastatic setting, the time until definitive deterioration, with or without death from all causes as event seems to be more adaptive, reflecting the absorbing state of health deterioration.[39]. Due to the multiplicity of the possible definitions of TTD (depending on the choice of the reference

score), MCID, and missing scores (including all-cause death or not), some recommendations have been proposed[44]. However, consensus guidelines by cancer localization, like RECIST criteria for HRQoL[44], are required to achieve standardization.[4] As already demonstrated for PFS[45], the time interval between assessments of HRQoL could influence Kaplan Meier estimation, resulting in an overestimation of TTD. Since the occurrence of the true time of HRQoL deterioration could be unknown, statistic approaches dealing with interval assessment may be proposed. These models are also based on the approximate definition of the MCID, which could be regarded as a limitation of this approach.

These models can also take into account death as an event, pending that for such a composite endpoint treatment must have the same effect on each component.[46] Effect size could be summarized with Hazard Ratio or restricted means.[47] The methodological challenge of such an approach would be to deal with the occurrence of competitive risks (between death and HRQoL score or between HRQoL scores), the interval measurement and the MCID definition. However, pending the use of the same definition, this approach allows comparison between trials.

As illustration of this survival approach, the randomized controlled BOLERO-2 (Breast Cancer Trials of Oral Everolimus) trial demonstrated significantly improved PFS with the use of everolimus plus exemestane versus placebo plus exemestane in patients with advanced breast cancer who developed disease progression after treatment with nonsteroidal aromatase inhibitors. The trial also used TDD to analyze longitudinal HRQoL data as a secondary endpoint[41]. This analysis included a protocol-specified TTD analysis at a 5% decrease in HRQOL versus baseline, with no subsequent increase above this threshold. The median TTD in global health status dimension was 8.3 months with everolimus plus exemestane versus 5.8 months with exemestane alone (hazard ratio, 0.74; P=0.0084). HRQoL results were

convergent to those obtained for PFS, which was the primary endpoint, thus enhancing the conclusion regarding the clinical benefit of this therapeutic strategy.

- Item response theory

Extension of IRT models to the longitudinal analysis is a recent statistical approach, but currently it is only explored for dichotomous items and three measurement times[48]. For polytomous items, IRT models the probability for one individual to choose a given response category for one item given the latent trait (i.e. HRQoL level) and the difficulty parameter of the item through a multinomial logit link function. Item difficulty parameter illustrates the difficulty to choose high response to the item. The modeling of the treatment effect depends on the conceptualization of the longitudinal model: it can correspond to a change in patient's latent trait level[49] or in item difficulty parameter [50]. A statistical challenge would be to extend and to validate such models for more than three measurement times and polytomous items. These models are potentially more adaptive than LMM or TTD approaches to the construction of questionnaires with few items per dimension, as for the major dimensions of the EORTC questionnaires' dimensions. Despite some interesting properties such as their robustness to missing data[51], IRT models do not produce clinically meaningful results for the clinicians.[52] These have not yet been applied to the longitudinal analysis of HRQoL in oncology clinical trials.

All these methods have their strengths and weaknesses, and comparison through simulation studies seems mandatory (Table 1). At the moment, no guidelines for the longitudinal analysis, which compromise comparison between trials, exist. For example, two recent phase III clinical trials investigating the impact of adding bevacizumab to the standard therapy of newly diagnosed glioblastomas have applied two different approaches (LMM and TTD) to analyze longitudinal HRQoL data. The results are divergent and compromise the conclusion

about clinical value of adding bevacizumab, since OS was not improved.[13,14] In the absence of guidelines for longitudinal analyses of HRQoL, it seems necessary to systematically apply at least two methods in order to facilitate the comparison of results between trials. Hence, it is proposed to systematically apply both TTD/TUDD and mixed model.

Two other challenging factors in HRQoL analyses are missing data and a response shift (RS) effect.

- Missing data

Missing data, considered as missing not at random[53], can bias the longitudinal analysis if it is not adequately taken into account.[54] Patients may drop out before the planned end of the study, resulting in the absence of any available HRQoL data after the patient's drop out (i.e. attrition). Moreover, drop out occurs generally due to a deterioration of patient health status or death. Patients may also be too tired to fill the questionnaire entirely at a specific measurement time. This induces the potential risk to select a subpopulation of patients with better HRQoL levels and with available HRQoL data. The consequence is a decrease of statistical power to highlight a difference between the two treatment arms. A great challenge is to prevent missing data, but also to adequately assess its patterns in case of its occurrence, which could influence choice of statistical analysis.

- Population data set for HRQoL analyses

The choice of a trial population to be analyzed is also a crucial aspect needing harmonization. Intent-to-treat (ITT) is the best way to prevent bias and allows comparison between studies. However, if patients do not have HRQoL data, a subset of ITT population needs to be defined. One proposition would be to use modified ITT population that includes all ITT patients with available HRQoL data at baseline. Sensitivity analyses should then be conducted after using multiple imputations for missing HRQoL scores/items at baseline.

- Response Shift

RS can be defined as "a change in the meaning of one's self-evaluation of a target construct as a result of: (a) a change in the respondent's internal standards of measurement (i.e. scale recalibration); (b) a change in the respondent's values (i.e. the importance of component domains constituting the target construct, or (c) a redefinition of the target construct (i.e. reconceptualization)"[55].

RS effect can negatively affect results of longitudinal studies by changing patients' expectations towards the disease and HRQoL over time.[55,56] Due to the adaptation of the treatment toxicities and disease, patients may not assess their HRQoL level with the same internal references at all measurement time points. Moreover, this could be different according to treatment modalities.[57] For example, one patient reporting a lower fatigue level after treatment as compared to baseline measurement (i.e. before treatment) could be due to a change in the patient's internal reference to assess fatigue intensity: the patient could have overestimated his fatigue level at baseline as compared to post-treatment assessment. This could influence patients' assessment of HRQoL level and interpretation of the impact of treatment. The possible occurrence of the RS effect reflects an importance of the choice of the reference score for qualification of a change such as deterioration. As evolution of RECIST criteria, the best response became reference of tumor measurement under treatment, the best previous score (inverse of the "nadir") may be considered as the reference instead of the baseline HRQoL score to qualify a deterioration in the context of RS[58,59 44].

Design of studies with HRQoL as endpoint

In cancer clinical trials, HRQoL is often inadequately captured. The EORTC recommends a minimum of three measurements during clinical trials: before, during, and after treatment.[5] A

more intensive longitudinal assessment is generally encouraged to ensure the capture of clinically important differences.

If HRQoL is a secondary endpoint in the trial, it is necessary to target the dimensions of interest (and to clearly specify them in the protocol) to avoid the risk of type I error inflation.

If HRQoL is the primary endpoint, the sample size is generally determined by comparing HRQoL at one follow-up measurement time to baseline HRQoL with a t-test or Wilcoxon non parametric test, in addition to target the HRQoL dimensions of interest. It would be more relevant to take into account the statistical method retained for longitudinal analysis. Sample size calculation is complicated because of probable time-by-treatment interaction which has to be taken into account in both LMM/IRT models. But the same limitations could be stressed for the survival analyses approach. For example, a time-by-treatment interaction is anticipated, then it would be more difficult to formulate hypotheses and to calculate sample size. In both cases, one has to specify the pattern and magnitude of that interaction, and it is often easier to think about it in terms of the original HRQoL scores rather than the time to achieving a MCID (or rather the hazard ratio for that event). Regarding LMM/IRT methods, the structure of the correlation must also be taken into account in the determination of the sample size. An alternative approach for trials design could be to incorporate HRQoL measures, as co-primary endpoints, with PFS or tumor response for example. Trials based on this approach may be declared successful, pending statistically significant treatment benefit on both with at least one dimension of HRQoL without deterioration of the other.

Testing multiple hypotheses without any adjustment of type I error (false positive) may increase the probability of erroneously rejecting at least one true null hypothesis. Consequently, according to the number of targeted HRQoL dimensions and/or the use of HRQoL and PFS as co-primary endpoints, the procedures to control the type I error must be clearly specified in the study protocol. For example, Hussain *et al.* used such design to

compare intermittent and continuous androgen deprivation in prostate cancer. In this study *P*-value of 0.01 was chosen for HRQoL to control the overall type I error rate at 0.05.[60] Therefore, the determination of sample size, taking into account the adjustment of the type I error and type of longitudinal analysis is an urgents statistical challenge to promote such designs.

Conclusion

Prolongation and improvement of HRQoL is a major endpoint that reflects a true clinical benefit for patients. Nevertheless, the lack of consensus on HRQoL definition, cutoff for MCID in HRQoL scores, guidelines for longitudinal HRQoL analysis, and results reporting are the main issues to overcome in future clinical trials so that such guidelines can be established to help clinicians in their decision making.

Table 1. Advantages and disadvantages of different methods for the longitudinal analysis of HRQoL data in oncology.

	Advantages	Disadvantages
linear mixed model for repeated measures	Well established approach	Response shift is not taken into account
	Easy to perform	Not easy to understand for the clinicians
	Robust to missing data	A sample size cannot be easily defined
Time to quality of life score deterioration	Clinically meaningful results	Need enough measures
	Missing data is easily taken into account	Not recommended for dimension evaluated by uni-item
	Death could be taken into account	Multiplicity of the existing definitions
	Response shift can be taken into account	Competitive risk has to be taken into account
		Interval measures has to be taken into account
		Based on the MCID definition which is very approximate
		A sample size cannot be easily defined
Item response theory	Robust to missing data	Could be less powerful for design integrating numerous HRQoL measures
	Deals with the latent measure of HRQoL	Difficult to understand, even for a statistician
	Relevant for uni-item dimension	Not implemented in main statistical software
		MCID has to be taken into account
		Response shift is no taken into account
		A sample size cannot be easily defined

HRQoL, Health related Quality of Life; MCID, minimal clinically important difference

1. Fiteni F, Westeel V, Pivot X, et al: Endpoints in cancer clinical trials. J Visc Surg 151:17-22, 2014

2. Blumenthal GM, Scher NS, Cortazar P, et al: First FDA approval of dual anti-HER2 regimen: pertuzumab in combination with trastuzumab and docetaxel for HER2-positive metastatic breast cancer. Clin Cancer Res 19:4911-6, 2013

3. Thornton K, Kim G, Maher VE, et al: Vandetanib for the treatment of symptomatic or progressive medullary thyroid cancer in patients with unresectable locally advanced or metastatic disease: U.S. Food and Drug Administration drug approval summary. Clin Cancer Res 18:3722-30, 2012

4. Bellera CA, Pulido M, Gourgou S, et al: Protocol of the Definition for the Assessment of Time-to-event Endpoints in CANcer trials (DATECAN) project: formal consensus method for the development of guidelines for standardised time-to-event endpoints' definitions in cancer clinical trials. Eur J Cancer 49:769-81, 2013

5. Aaronson NK, Ahmedzai S, Bergman B, et al: The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 85:365-76, 1993

6. Cella DF, Tulsky DS, Gray G, et al: The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol 11:570-9, 1993

7. Sprangers MA, Cull A, Bjordal K, et al: The European Organization for Research and Treatment of Cancer. Approach to quality of life assessment: guidelines for developing questionnaire modules. EORTC Study Group on Quality of Life. Qual Life Res 2:287-95, 1993

8. Sprangers MA, Groenvold M, Arraras JI, et al: The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: first results from a three-country field study. J Clin Oncol 14:2756-68, 1996

9. Bergman B, Aaronson NK, Ahmedzai S, et al: The EORTC QLQ-LC13: a modular supplement to the EORTC Core Quality of Life Questionnaire (QLQ-C30) for use in lung cancer clinical trials. EORTC Study Group on Quality of Life. Eur J Cancer 30A:635-42, 1994

10. Brady MJ, Cella DF, Mo F, et al: Reliability and validity of the Functional Assessment of Cancer Therapy-Breast quality-of-life instrument. J Clin Oncol 15:974-86, 1997

11. Cella DF, Bonomi AE, Lloyd SR, et al: Reliability and validity of the Functional Assessment of Cancer Therapy-Lung (FACT-L) quality of life instrument. Lung Cancer 12:199-220, 1995

12. Edelen MO, Reeve BB: Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 16 Suppl 1:5-18, 2007

13. Chinot OL, Wick W, Mason W, et al: Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. N Engl J Med 370:709-22, 2014

14. Gilbert MR, Dignam JJ, Armstrong TS, et al: A randomized trial of bevacizumab for newly diagnosed glioblastoma. N Engl J Med 370:699-708, 2014

15. Chen G, Feng J, Zhou C, et al: Quality of life (QoL) analyses from OPTIMAL (CTONG-0802), a phase III, randomised, open-label study of first-line erlotinib versus chemotherapy in patients with advanced EGFR mutation-positive non-small-cell lung cancer (NSCLC). Ann Oncol 24:1615-22, 2013

16. Yang JC, Hirsh V, Schuler M, et al: Symptom Control and Quality of Life in LUX-Lung 3: A Phase III Study of Afatinib or Cisplatin/Pemetrexed in Patients With Advanced Lung Adenocarcinoma With EGFR Mutations. J Clin Oncol, 2013

17. Postel-Vinay S, Arkenau HT, Olmos D, et al: Clinical benefit in Phase-I trials of novel molecularly targeted agents: does dose matter? Br J Cancer 100:1373-8, 2009

18. Bruner DW, Hanisch LJ, Reeve BB, et al: Stakeholder perspectives on implementing the National Cancer Institute's patient-reported outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). Transl Behav Med 1:110-22, 2011

19. Hay JL, Atkinson TM, Reeve BB, et al: Cognitive interviewing of the US National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). Qual Life Res 23:257-69, 2014

20. Osoba D: Health-related quality of life and cancer clinical trials. Ther Adv Med Oncol 3:57-71, 2011

21. Fuchs CS, Tomasek J, Yong CJ, et al: Ramucirumab monotherapy for previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (REGARD): an international, randomised, multicentre, placebo-controlled, phase 3 trial. Lancet 383:31-9, 2014

22. Fairclough DL: Design and analysis of quality of life studies in clinical trials, CRC press, 2010

23. Jaeschke R, Singer J, Guyatt GH: Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 10:407-15, 1989

24. Osoba D, Rodrigues G, Myles J, et al: Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 16:139-44, 1998

25. Cocks K, King MT, Velikova G, et al: Evidence-based guidelines for determination of sample size and interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. J Clin Oncol 29:89-96, 2011

26. Maringwa J, Quinten C, King M, et al: Minimal clinically meaningful differences for the EORTC QLQ-C30 and EORTC QLQ-BN20 scales in brain cancer patients. Ann Oncol 22:2107-12, 2011

27. Maringwa JT, Quinten C, King M, et al: Minimal important differences for interpreting health-related quality of life scores from the EORTC QLQ-C30 in lung cancer patients participating in randomized controlled trials. Support Care Cancer 19:1753-60, 2011

28. Cella D, Eton DT, Fairclough DL, et al: What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol 55:285-95, 2002

29. Cella D, Eton DT, Lai JS, et al: Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. J Pain Symptom Manage 24:547-61, 2002

30. Eton DT, Cella D, Yost KJ, et al: A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol 57:898-910, 2004

31. Terwee CB, Roorda LD, Dekker J, et al: Mind the MIC: large variation among populations and methods. J Clin Epidemiol 63:524-34, 2010

32. Norman GR, Sloan JA, Wyrwich KW: Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 41:582-92, 2003

33. Cnaan A, Laird NM, Slasor P: Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Stat Med 16:2349-80, 1997

34. Little RJ, Wang Y: Pattern-mixture models for multivariate incomplete data with covariates. Biometrics 52:98-111, 1996

35. Stockler MR, Hilpert F, Friedlander M, et al: Patient-reported outcome results from the open-label phase III AURELIA trial evaluating bevacizumab-containing therapy for platinum-resistant ovarian cancer. J Clin Oncol 32:1309-16, 2014

36. Brundage M, Osoba D, Bezjak A, et al: Lessons learned in the assessment of health-related quality of life: selected examples from the National Cancer Institute of Canada Clinical Trials Group. J Clin Oncol 25:5078-81, 2007

37. Osoba D, Bezjak A, Brundage M, et al: Analysis and interpretation of health-related quality-of-life data from clinical trials: basic approach of The National Cancer Institute of Canada Clinical Trials Group. Eur J Cancer 41:280-7, 2005

38. Calvert M, Blazeby J, Altman DG, et al: Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA 309:814-22, 2013

39. Bonnetain F, Dahan L, Maillard E, et al: Time until definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. Eur J Cancer 46:2753-62, 2010

40. Gourgou-Bourgade S, Bascoul-Mollevi C, Desseigne F, et al: Impact of FOLFIRINOX compared with gemcitabine on quality of life in patients with metastatic pancreatic cancer: results from the PRODIGE 4/ACCORD 11 randomized trial. J Clin Oncol 31:23-9, 2013

41. Burris HA, 3rd, Lebrun F, Rugo HS, et al: Health-related quality of life of patients with advanced breast cancer treated with everolimus plus exemestane versus placebo

plus exemestane in the phase 3, randomized, controlled, BOLERO-2 trial. Cancer 119:1908-15, 2013

42. Kabbinavar FF, Wallace JF, Holmgren E, et al: Health-related quality of life impact of bevacizumab when combined with irinotecan, 5-fluorouracil, and leucovorin or 5-fluorouracil and leucovorin for metastatic colorectal cancer. Oncologist 13:1021-9, 2008

43. Hamidou Z, Dabakuyo TS, Mercier M, et al: Time to deterioration in quality of life score as a modality of longitudinal analysis in patients with breast cancer. Oncologist 16:1458-68, 2011

44. Anota A, Hamidou Z, Paget-Bailly S, et al: Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization? Quality of Life Research, 2013

45. Panageas KS, Ben-Porat L, Dickler MN, et al: When you look matters: the effect of assessment schedule on progression-free survival. J Natl Cancer Inst 99:428-32, 2007

46. Freemantle N, Calvert M, Wood J, et al: Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA 289:2554-9, 2003

47. Royston P, Parmar MK: Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol 13:152, 2013

48. Blanchin M, Hardouin JB, Le Neel T, et al: Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes. Stat Med 30:825-38, 2011

49. Bacci S: Analysis of longitudinal HrQoL using latent regression in the context of Rasch modeling. Mathematical Methods in Survival Analysis, Reliability and Quality of Life:275-290, 2008

50. Fischer GH, Ponocny I: An extension of the patial credit model with an application to the measurement of change. Psychometrika 59:177-192, 1994

51. De Ayala RJ: The theory and practice of item response theory. New York : Guilford Press,. 2009

52. Rouquette A, Blanchin M, Sebille V, et al: The minimal clinically important difference determined using item response theory models: an attempt to solve the issue of the association with baseline score. J Clin Epidemiol 67:433-40, 2014

53. Little RJ, Rubin DB: Statistical analysis with missing data. New York: John Wiley & Sons, 1987

54. Bernhard J, Cella DF, Coates AS, et al: Missing quality of life data in cancer clinical trials: serious problems and challenges. Stat Med 17:517-32, 1998

55. Sprangers MA, Schwartz CE: Integrating response shift into health-related quality of life research: a theoretical model. Soc Sci Med 48:1507-15, 1999

56. Schwartz CE, Sprangers MA: Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc Sci Med 48:1531-48, 1999

57. Ring L, Hofer S, Heuston F, et al: Response shift masks the treatment impact on patient reported outcomes (PROs): the example of individual quality of life in edentulous patients. Health Qual Life Outcomes 3:55, 2005

58. Eisenhauer EA, Therasse P, Bogaerts J, et al: New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 45:228-47, 2009

59. Therasse P, Arbuck SG, Eisenhauer EA, et al: New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 92:205-16, 2000

60. Hussain M, Tangen CM, Berry DL, et al: Intermittent versus continuous androgen deprivation in prostate cancer. N Engl J Med 368:1314-25, 2013

2. Temps jusqu'à détérioration d'un score de QdV comme modalité d'analyse longitudinale de la QdV en Cancérologie

2.1. Proposition de recommandations pour l'analyse longitudinale selon la méthode du temps jusqu'à détérioration d'un score de QdV

Résumé

Contexte

L'analyse longitudinale de la QdV reste complexe et non standardisée ce qui compromet la comparaison des résultats des essais cliniques. En cancérologie, les résultats sont encore peu utilisés pour changer les standards de prises en charge des patients en raison d'un manque de standardisation de l'analyse longitudinale et d'une incapacité de ces méthodes statistiques à proposer des résultats qui soit compréhensibles par les cliniciens.

Dans cette optique, le temps jusqu'à détérioration a été proposé comme modalité d'analyse longitudinale en cancérologie. En raison de la multiplicité des définitions possibles, des recommandations sur les définitions à utiliser semblent nécessaires, tels que des critères RECIST (« Response Evaluation Criteria In Solid Tumors ») pour la QdV, tenant compte de la situation thérapeutique et de l'occurrence potentielle d'un effet Response Shift.

Matériels et méthodes

Différentes définitions de TJD ont été explorées et présentées dans cet article. En situation adjuvante, il parait pertinent d'utiliser le TJD simple en tant qu'état transitoire tandis qu'en situation avancée, l'utilisation du TJD en tant qu'état définitif ou récurrent selon la définition de Bonnetain et al. (Bonnetain *et al*, 2010), intégrant ou non le décès dans la définition de l'évènement, semble plus adaptée. De plus, le choix du score de référence pour qualifier la détérioration dépend de l'occurrence ou non d'un effet Response Shift démontré. Ainsi, en absence d'effet Response Shift, le score à l'inclusion peut être considéré comme le score de référence. Si un effet Response Shift est démontré, le meilleur score antérieur (inverse du Nadir) ou le dernier score immédiatement précédent peuvent être considérés comme score de référence.

Ces approches ont été appliquées respectivement à une étude portant sur le cancer du sein en situation adjuvante ainsi qu'à un essai clinique de phase II portant sur le cancer du pancréas en situation métastatique avec deux bras de traitement (Gemcitabine seule vs. FOLFIRI.3 + gemcitabine). Dans l'étude portant sur le cancer du sein, le TJD a été défini par rapport au score à l'inclusion puis par rapport au meilleur score antérieur puisqu'un effet Response Shift avait déjà été démontré. Concernant l'étude sur le cancer du pancréas métastatique, le TUDD a été défini selon la définition de Bonnetain et al. (Bonnetain *et al*, 2010). Trois définitions ont été appliquées :

1. TUDD d'au moins 5 points par rapport à l'inclusion ;
2. En considérant le décès comme évènement ;
3. En considérant les patients sans mesure de suivi en détérioration dès l'inclusion.

Résultats

381 femmes ont été incluses dans l'étude portant sur le cancer du sein. La médiane de détérioration était influencée par le choix du score de référence. Pour les symptômes au niveau du bras et du sein, la médiane de détérioration augmentait lorsque le meilleur score antérieur était choisi comme score de référence au lieu du score à l'inclusion:

- de 2.9 mois [IC 95% 0.4 – 3.1] à 6.0 mois [IC 95% 3.6 – 6.0] pour les symptômes au niveau du bras ;
- de 0.2 mois [IC 95% 0.2 – 2.8] à 2.8 mois [IC 95% 2.8 – 3.0] pour les symptômes au niveau du sein.

98 patients ont été inclus dans l'essai de phase II portant sur le cancer du pancréas métastatique. Les patients du bras FOLFIRI.3 + gemcitabine présentaient un TJDD globalement plus long que ceux du bras de référence (gemcitabine seule), quel que soit la définition appliquée.

Concernant les définitions intégrant ou non le décès comme évènement, les patients du bras FOLFIRI.3 + gemcitabine présentaient une détérioration significativement plus tardive que ceux du bras gemcitabine seule pour le fonctionnement physique. Cette tendance n'était plus significative lorsque les patients sans mesure de suivi étaient considérés en détérioration.

Conclusion

Le TJD est une approche didactique et prometteuse qui peut être recommandée comme modalité d'analyse longitudinale de la QdV en cancérologie. En particulier, cette méthode permet de tenir compte et de s'adapter aux différentes situations thérapeutiques ainsi qu'à l'occurrence d'un effet Response Shift.

<u>Article:</u> Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization?

Article accepté dans *Quality of Life Research*

Qual Life Res
DOI 10.1007/s11136-013-0583-6

QUANTITATIVE METHODS SPECIAL SECTION

Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization?

Amélie Anota · Zeinab Hamidou · Sophie Paget-Bailly · Benoist Chibaudel · Caroline Bascoul-Mollevi · Pascal Auquier · Virginie Westeel · Frederic Fiteni · Christophe Borg · Franck Bonnetain

Accepted: 12 November 2013

Abstract
Purpose Longitudinal analysis of health-related quality of life (HRQoL) remains unstandardized and compromises comparison of results between trials. In oncology, despite available statistical approaches, results are poorly used to change standards of care, mainly due to lack of standardization and the ability to propose clinical meaningful results. In this context, the time to deterioration (TTD) has been proposed as a modality of longitudinal HRQoL analysis for cancer patients. As for tumor response and progression, we propose to develop RECIST criteria for HRQoL.
Methods Several definitions of TTD are investigated in this paper. We applied this approach in early breast cancer and metastatic pancreatic cancer with a 5-point minimal clinically important difference. In breast cancer, TTD was defined as compared to the baseline score or to the best previous score. In pancreatic cancer (arm 1: gemcitabine with FOLFIRI.3, arm 2: gemcitabine alone), the time until definitive deterioration (TUDD) was investigated with or without death as event.
Results In the breast cancer study, 381 women were included. The median TTD was influenced by the choice of the reference score. In pancreatic cancer study, 98 patients were enrolled. Patients in Arm 1 presented longer TUDD than those in Arm 2 for most of HRQoL scores. Results of TUDD were slightly different according to the definition of deterioration applied.
Conclusion Currently, the international ARCAD group supports the idea of developing RECIST for HRQoL in pancreatic and colorectal cancer with liver metastasis, with a view to using HRQoL as a co-primary endpoint along with a tumor parameter.

Keywords Health-related quality of life · Clinical trials · Oncology · Longitudinal analysis · Time to deterioration

Electronic supplementary material The online version of this article (doi:10.1007/s11136-013-0583-6) contains supplementary material, which is available to authorized users.

A. Anota · Z. Hamidou · P. Auquier · F. Bonnetain
Quality of Life in Oncology Clinical Research Platform, Besançon, France

A. Anota (✉) · S. Paget-Bailly · F. Fiteni · F. Bonnetain
Methodological and Quality of Life in Oncology Unit, EA 3181, University Hospital of Besançon, 2 Place Saint-Jacques, 25030 Besançon Cedex, France
e-mail: aanota@chu-besancon.fr

Z. Hamidou · P. Auquier
Public Health Laboratory, EA 3279, Aix-Marseille University, Marseille, France

B. Chibaudel
Medical Oncology Department, University Hospital Saint-Antoine, Paris, France

B. Chibaudel
Gercor, Clinical Research Group in Oncology, Paris, France

C. Bascoul-Mollevi
Department of Biostatistics, Regional Cancer Institute, Montpellier, France

V. Westeel
Pneumology Department, University Hospital of Besançon, Besançon, France

F. Fiteni · C. Borg
Medical Oncology Department, University Hospital of Besançon, Besançon, France

Published online: 26 November 2013

Abbreviations

CI	Confidence interval
EORTC	European Organisation for Research and Treatment of Cancer
HR	Hazard ratio
HRQoL	Health-related quality of life
GLMM	General linear mixed model
IRT	Item response theory
MCID	Minimal clinically important difference
OS	Overall survival
RS	Response shift
SD	Standard deviation
TTD	Time to deterioration
TUDD	Time until definitive deterioration

Introduction

Although overall survival (OS) is still considered as the "gold standard" for primary endpoints in many oncology studies, most clinical trials now integrate health-related quality of life (HRQoL) as one of the major endpoints to investigate the clinical benefit of new therapeutic strategies for the patient. HRQoL is considered as a second primary endpoint by the American Society of Clinical Oncology and the Food and Drug Administration if no effect of treatment on OS is observed [1–3]. Moreover, since many trials in oncology use so-called surrogate endpoints for OS focusing on tumor parameters, it is of major importance to assess HRQoL in order to characterize the clinical benefit for patients.

Despite this opportunity to achieve comprehensive assessment of HRQoL to support "evidence-based medicine" in oncology, the longitudinal analysis of HRQoL remains unstandardized. This compromises the comparison of results between trials. Moreover, longitudinal results should translate findings into information that decision-makers find understandable and compelling. However, despite the many sophisticated statistical approaches available, results remain underutilized in clinical practice, especially due to a lack of standardization and the inability to propose clinically meaningful results.

Analyses also have to deal with another limiting factor, namely missing data. Patients may not complete the entire HRQoL questionnaire at all planned measurement times. Moreover, patients may drop out before the end of the study, generally due to a deterioration of their health status, or death, as in the palliative setting. Missing data can bias the analysis and interpretation of the results if they depend on the patient's health status [4–6]. Therefore, there is a need to develop statistical methods that can handle missing data [7–12].

Another challenge of longitudinal HRQoL analysis is to take into account the potential occurrence of a response shift (RS) effect. Indeed, self-assessment of HRQoL is dependent on the patient's internal standards and the definition of HRQoL used [13–15]. Since patients can adapt to disease and the treatment toxicities, their health and HRQoL expectations may also change over time. These changes result in an RS effect [16]. Sprangers and Schwartz defined RS as "a change in the meaning of one's self-evaluation of a target construct as a result of the following: (a) a change in the respondent's internal standards of measurement (i.e., scale recalibration); (b) a change in the respondent's values (i.e., the importance of component domains constituting the target construct); or (c) a redefinition of the target construct (i.e., reconceptualization)" [17]. Thus, the choice of the reference score to qualify a change such as deterioration is a major concern.

Several methods are used to analyze longitudinal HRQoL data [18–20]. The most widely used is the general linear mixed model (GLMM) [18, 21–23], which is recommended in longitudinal studies with a limited number of follow-up [24]. This method is only adapted when HRQoL assessments are widely spaced and with little amplitude within patients. GLMM can handle the missing data profiles by applying a pattern mixture model [10, 25]. However, these sub-models are rarely applied, mainly because of the complexity of the pattern construction [10, 25–27]. Furthermore, GLMM does not deal with the occurrence of a RS effect.

In the last few years, researchers have started to use models of modern item response theory (IRT) to analyze longitudinal HRQoL data [28]. In contrast to the GLMM, the link between the observed score and the latent trait (e.g., HRQoL) is not linear but logistic. However, these models are rarely used to analyze longitudinal HRQoL data, mainly due to their complexity [29].

Also in recent years, time-to-event models such as the time-to-HRQoL score deterioration (TTD) have been proposed as an approach to the analysis of longitudinal HRQoL in oncology [30, 31]. Both GLMM and TTD rely on the definition of the minimal clinically important difference (MCID) in order to be effective from a clinical point of view. The measure of TTD might be more familiar to clinicians because it is based on Kaplan–Meier survival curves and hazard ratios (HR). As for GLMM, TTD can deal with missing data by making underlying assumptions about whether the missing data reflect a deterioration of the patient's health status or not. Contrary to GLMM, the TTD method can take into account the occurrence of the RS recalibration component by choosing different reference scores to qualify the deterioration.

TTD cannot be considered as an exclusive method, since the GLMM approach measures different concepts and

proposes complementary ways of summarizing HRQoL data. However, if few HRQoL assessments are performed and the interval time between two consecutive assessments is long, then GLMM may be more relevant than the TTD approach. In other cases, the TTD approach may be more suitable than GLMM.

Regarding the TTD approach, the choice of event definition is essential, because it may lead to different results. However, there are currently no recommendations or consensus in this regard, with the result that TTD reflects heterogeneity.

Thus, there is a clear need to investigate and validate several definitions of TTD depending on the following: the cancer context (adjuvant, advanced), reference score, event definitions, MCID, and censoring rules. As for tumor response and progression, one proposition could be to develop "RECIST" criteria ("Response Evaluation Criteria In Solid Tumors") for HRQoL. This would allow standardization of longitudinal HRQoL analysis using the TTD method, according to the therapeutic situation and the cancer site. Accordingly, several definitions of TTD were investigated and are presented in this paper. We next propose recommendations for the choice of the definition depending on the therapeutic situation. Finally, we report results observed using the TTD approach in early breast cancer and metastatic pancreatic cancer.

Methods

Time to deterioration definitions

We propose several definitions of TTD in a HRQoL score according to the therapeutic situation and cancer site. Events can be defined in relation to a reference score, MCID, and missing scores, including death or not. These definitions are summarized in Table 1.

1) Core definitions with respect to the MCID

The most intuitive definition for TTD is the time from inclusion–randomization in the study to

- a first deterioration of at least one MCID unit as compared to the baseline score [31] (Fig. 1a).
- Patients with no deterioration before their dropout are censored at the time of the last follow-up or the last HRQoL assessment.

This definition corresponds to definition TTD#1 in Table 1.

According to the scoring algorithm of the HRQoL dimension, the deterioration corresponds to an increase or decrease in at least one MCID unit of the score as compared to the baseline score. The MCID may vary depending on the instruments and cancer sites under consideration.

The deterioration observed can be definitive or not. In the palliative setting, Bonnetain et al. have previously defined the time until definitive HRQoL score deterioration (TUDD) as the time from inclusion in the study to a first deterioration of at least one MCID unit as compared to the baseline score:

- with no further improvement of more than one MCID unit as compared to the baseline score (Fig. 1b).
- or if the patient dropped out after deterioration, resulting in missing data.

This corresponds to the definition TUDD#1 in Table 1.

An alternative for defining TUDD is to consider that the first deterioration of at least one MCID unit observed at time T is definitive:

- if the deterioration of at least one MCID unit as compared to the baseline score is also observed at all time points after time T (Fig. 1c).
- or if the patient dropped out after deterioration, resulting in missing data.

This second definition of TUDD corresponds to definition TUDD#5 of Table 1.

2) Alternatives for defining the reference score

The concept of deterioration requires a reference score relative to which the deterioration may be quantified. In the definitions described here, the reference score is the baseline score. However, the reference score could also be defined in other ways. For example,

- the best previous HRQoL score. Figure 1d illustrates the TTD with a 10-point MCID as compared to the best previous HRQoL score for one patient (TTD#5 in Table 1) or
- the previous HRQoL score. Figure 1e illustrates the TTD with a 10-point MCID for one patient with the previous score (i.e., "immediately preceding score") as the reference score (TTD#9 in Table 1).

Moreover, for definitive deterioration, the deterioration observed at time T can be considered definitive:

- as compared to the reference score (baseline score, previous score, or best previous score) or
- as compared to the score qualifying the deterioration (i.e., the score obtained at time T). In that case, the score qualifying the deterioration at time T becomes the reference score (TUDD#9). Figure 1f. illustrates the TUDD as compared to the baseline score with no

Table 1 Summary of the different definitions of time to deterioration (TTD) and time until definitive HRQoL score deterioration (TUDD) investigated

To be considered as events	Reference score			Definitive as compared to			Death	Patients with no baseline	Patients with no follow-up
	Baseline	Best previous score	Previous score	Reference score		Score qualifying the deterioration			
				MCID+[a]	MCID−[b]				
TTD									
1	X								
2	X							X	X
3	X						X		
4	X						X	X	X
5		X							
6		X						X	X
7		X					X		
8		X					X	X	X
9			X						
10			X					X	X
11			X				X		
12			X				X	X	X
TUDD									
1	X			X					
2	X			X				X	X
3	X			X			X		
4	X			X			X	X	X
5	X				X				
6	X				X			X	X
7	X				X		X		
8	X				X		X	X	X
9	X					X			
10	X					X		X	X
11	X					X	X		
12	X					X	X	X	X
13		X		X					
14		X		X				X	X
15		X		X			X		
16		X		X			X	X	X
17		X			X				
18		X			X			X	X
19		X			X		X		
20		X			X		X	X	X
21		X				X			
22		X				X		X	X
23		X				X	X		
24		X				X	X	X	X
25			X	X					
26			X	X				X	X
27			X	X			X		
28			X	X			X	X	X
29			X		X				
30			X		X			X	X
31			X		X		X		

Table 1 continued

To be considered as events	Reference score			Definitive as compared to			Death	Patients with no baseline	Patients with no follow-up
	Baseline	Best previous score	Previous score	Reference score		Score qualifying the deterioration			
				MCID+[a]	MCID−[b]				
32			X		X		X	X	X
33			X			X			
34			X			X		X	X
35			X			X	X		
36			X			X	X	X	X

A cross (X) indicates the retained definition and the corresponding events

[a] MCID+ deterioration with no further improvement as compared to the reference score (definition of Bonnetain et al.)

[b] MCID− definitive deterioration if deterioration observed at all time points following the initial deterioration

further improvement as compared to the score qualifying the deterioration for one patient.

3) Missing data issues

Intermittent missing data are ignored in the TTD approach, which goes on the assumption that HRQoL level remains unchanged since the last available HRQoL assessment. Moreover, patients with no baseline HRQoL score or with no follow-up score are usually excluded from longitudinal analysis. However, these patients can be included in the analysis and censored at baseline or just after baseline. Depending on the therapeutic situation, sensitivity analysis can be performed considering these patients to be deteriorating since baseline. For example, definition TUDD#2 in Table 1 corresponds to TUDD as compared to the baseline score, according to the definition of Bonnetain et al., including patients with no baseline HRQoL score or with no follow-up score as events.

4) Death as an event

All-cause death can be considered as an event if the patient did not experience deterioration before death. These supplementary events (death, no follow-up) will be addressed in the case of TUDD. In this way, TUDD or death could be redefined as "HRQoL deterioration-free survival." For example, definition TUDD#3 in Table 1 corresponds to TUDD as compared to the baseline score according to the definition of Bonnetain et al., or death.

5) Response shift issue

Patients' internal standards can change over time, reflecting the recalibration component of RS. An alternative way to take into account the occurrence of the recalibration component of RS could be to consider the reference score as the best previous HRQoL score, or the previous (immediately preceding) HRQoL score but not the baseline sore. The value of these scores can change over time according to the patient's experience of treatment and disease course.

6) Multidimensional definition

We can study the deterioration of one given HRQoL score, or the deterioration of at least one HRQoL dimension among the set of all dimensions. For example, we can study deterioration of at least one dimension of a multidimensional questionnaire. In the case of a multidimensional definition, the event time corresponds to the first deterioration observed, irrespective of which HRQoL score is affected. In this situation, competitive risks should be taken into account. This multidimensional definition has the advantage of increasing the statistical power and may be relevant if the treatment is expected to have a similar effect on all the HRQoL dimensions retained.

As TTD analyses count as survival analyses, the TTD estimation can be calculated using the Kaplan–Meier or actuarial method and described using median and 95 % confidence interval (CI). The Kaplan–Meier survival curve is defined as the probability of surviving in a given length of time while considering time in many small intervals. This method is based on the intuitive idea that being alive at time T naturally requires the subject to be alive just before time T, and not to die at time T [32]. Contrary to the Kaplan–Meier method, in the actuarial method, probabilities are estimated for fixed time intervals, not determined by the date of observed death. Both methods can handle the presence of censored data, i.e., patients are still alive at the end of the study.

In time to deterioration (TTD) analyses, the event is "the HRQoL score deterioration." The Kaplan–Meier estimation is given by the following formula:

$$S(t) = \prod_{t_i \leq t} \frac{n_i - m_i}{n_i}$$

where $n_i = n_{i-1} - m_{i-1} - c_{i-1}$ and n_i is the number of subject at risk at time T_i, i.e., the number of patients still in the study and who do not present a deterioration until time

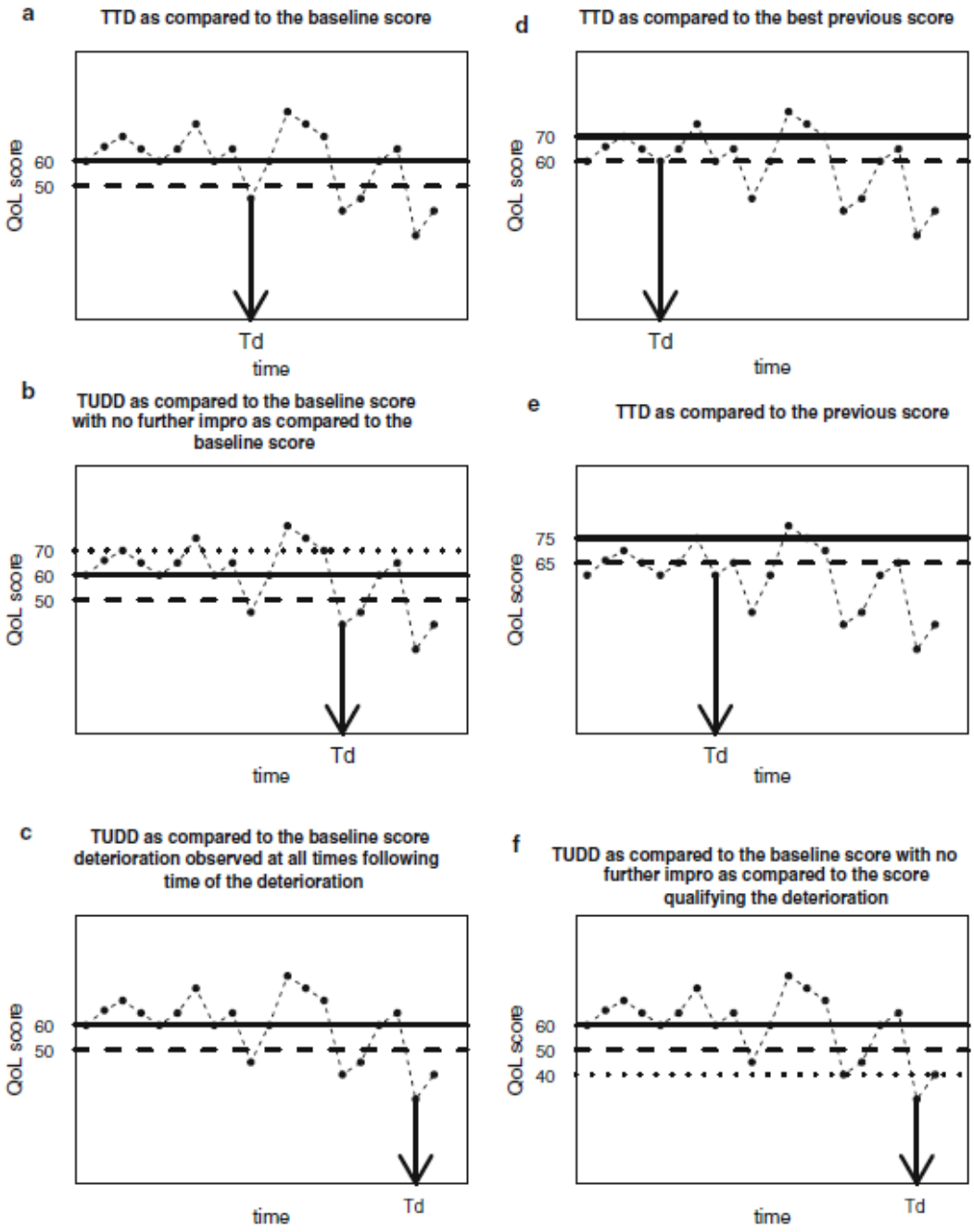


Fig. 1 Illustration of time to deterioration (Td) using different definitions with a 10-point MCID for one patient and for a health-related quality of life score (QoL) in which a deterioration corresponds to a decrease in the score. The *solid line* corresponds to the value of the reference score at time Td. The *dashed line* corresponds to the threshold to observe deterioration as compared to the reference score at time Td. The *dotted line* corresponds to the threshold to observe a definitive deterioration as compared to the reference score at time Td

T_{i-1}, m_i is the number of events observed at time T_i, i.e., the number of patients experiencing a HRQoL score deterioration at time T_i, and c_i is the number of censored patients at time T_i, i.e., the number of patients who dropped out at time T_i and who did not experienced a HRQoL deterioration before.

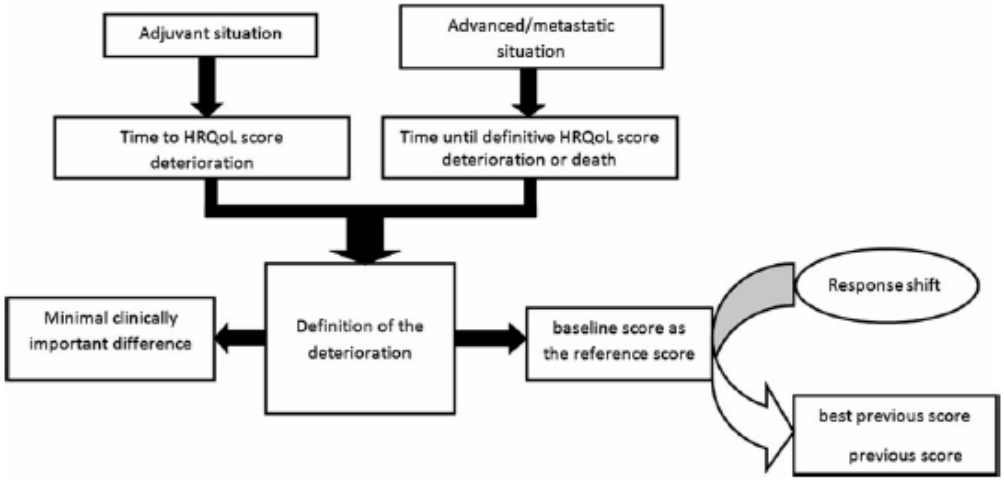


Fig. 2 Decision-making flowchart according to the therapeutic situation

TTD can then be compared according to treatment arm using the log-rank test and univariate Cox analyses to produce a HR with 95 % CI. Multivariate Cox regression can be applied to identify independent factors associated with TTD.

In Fig. 2, we propose a decision-making flowchart. In the adjuvant setting, we recommend using the TTD; and in the advanced or metastatic setting, we recommend using the TUDD, with or without death from all causes as an event. Indeed, it is intuitive that in the adjuvant setting, deterioration is expected not to be definitive, because the patient could conceivably survive the cancer. Moreover, cancer survivors can experience an improvement of their HRQoL. In contrast, in the advanced or metastatic setting, a definitive deterioration is more relevant, reflecting the deterioration of the patient's health status, which is stable over time. Furthermore, the time between deterioration and death is often short for these patients [30]. The definition of the deterioration is based on both the threshold for the MCID, and the definition chosen for the reference score. Thus, if no RS effect occurs, the baseline score can be kept as the reference score in the TTD analysis. If a RS is likely to occur, we recommend using the best previous score or the previous score as the reference score in the TTD analysis.

Health-related quality of life studies

In this section, we report TTD analyses performed in two studies as an illustration, namely early breast cancer and metastatic pancreatic cancer. In the breast cancer study, since it is an adjuvant setting, we retained the TTD approach and studied the impact of RS on TTD using changing score as the reference score, i.e., the best previous score. In the metastatic pancreatic cancer study, as it is a metastatic setting, we retained the TUDD approach, integrating death (or not) as event. We also took into account informative missing data.

Time to deterioration in early breast cancer

A prospective, multicenter, randomized, cohort study including all women hospitalized for the diagnosis or treatment of first primary breast cancer or for a suspicion of breast cancer was performed in French hospitals between February 2006 and February 2008. All participants gave written informed consent, and the local ethics committee approved the study protocol. The complete design of this study has previously been described elsewhere [33].

HRQoL was evaluated using the EORTC cancer-specific questionnaire QLQ-C30 [34] and its breast cancer module QLQ-BR23 [35]. These were administered at inclusion, at discharge following initial hospitalization, as well as at three and 6 months after inclusion. The QLQ-C30 and its breast cancer module BR23 are validated tools to assess HRQoL in cancer, specifically in breast cancer [34, 35].

The QLQ-C30 includes 30 items and measures five functional scales (physical, role, emotional, cognitive, and social functioning), global health status (GHS), financial difficulties, and eight symptom scales (fatigue, nausea and vomiting, pain, dyspnea, insomnia, appetite loss, constipation, and diarrhea) [34].

The BR23 module includes 23 items that generate four functional scales (body image, sexual functioning, sexual enjoyment, and future perspective) and four symptom scales (systemic therapy side effects, breast symptoms, arm symptoms, and upset caused by hair loss) [35].

The occurrence of a RS effect has already been demonstrated in early breast cancer patients [33, 36] and particularly in this study [31, 33]. Thus, two definitions of TTD were investigated using a 5-point MCID: The first definition was TTD with the baseline score as the reference score [31]. The second was TTD with the best previous score as the reference score. Patients with at least one HRQoL score were included in the TTD analysis. Patients with no follow-up HRQoL score were censored just after baseline. Patients with no baseline score were censored at baseline.

TTD curves were calculated using the Kaplan–Meier estimation and described using median and 95 % CI.

Time until definitive deterioration in metastatic pancreatic cancer

This study was a multicenter, randomized, open phase II trial conducted in 11 French centers between October 2007 and May 2011. Randomization 1:1 was done using the minimization technique with stratification according to center, performance status (0 vs. 1), and the number of metastatic sites (one vs. more than one).

Inclusion criteria were as follows: histologically or cytologically proven metastatic pancreatic adenocarcinoma, no previous chemotherapy, no previous radiotherapy, and WHO performance status <2.

Exclusion criteria were bile duct adenocarcinoma, ampulloma, and history of another major cancer.

All patients were fully informed of the study and provided written informed consent. The protocol was approved by the ethics committee.

Patients were randomly assigned to receive alternately FOLFIRI 3 every 14 days for 2 months (i.e., 4 courses per cycle), followed by gemcitabine, 6 courses at days 1, 8, 15, 29, 36, and 43 per cycle (Arm 1) or gemcitabine alone (Arm 2). FOLFIRI 3 is a chemotherapeutic regimen combining 5-fluorouracil, folinic acid, and irinotecan.

HRQoL was evaluated using the QLQ-C30 questionnaire [34] at inclusion and every 2 months until the end of the study or death.

The TUDD was defined as the TUDD with a 5-point MCID as compared to the baseline score, with no further improvement of more than 5 points [30]. Patients with at least one HRQoL score were included in the TUDD analysis. Patients with no baseline score were censored at baseline. Patients with no follow-up measures were censored just after baseline. Sensitivity analyses were conducted, first considering death as an event and then simultaneously considering death and no follow-up as events. TUDD analyses including death as an event are referred to "HRQoL deterioration-free survival" analyses.

TUDD curves were calculated using the Kaplan–Meier method and described using median and 95 % CI. TUDD was compared between treatment arms using the log-rank test and univariate HR with 95 % CI.

For both studies, variables collected at baseline are described as means and standard deviations (SD) for continuous variables and number (percentage) for qualitative variables. The percentage of missing data is also provided. The number of HRQoL questionnaires completed at each measurement time is reported. Scores were generated according to the EORTC scoring manual [37]. These scores vary from 0 (worst) to 100 (best) for the functional dimensions and GHS, and from 0 (best) to 100 (worst) for the symptom dimensions.

All analyses were performed with R software [38].

Results

Breast cancer

Between February 2006 and February 2008, 381 patients were included in the four participating centers. Mean age was 58.4 (SD = 11) years. Complete clinical and pathologic characteristics of the population are given in supplementary Table A.

At baseline, 359 (94 %) patients had at least one HRQoL score, 343 (90 %) at discharge following initial hospitalization, 340 (89 %) at three months, and 321 (84 %) at 6 months.

Results of the TTD analyses are summarized in Table 2. Among the 377 patients included with at least one cognitive functioning score, 160 and 197 patients presented deterioration of cognitive function as compared to the baseline score and the best previous score, respectively. The median TTD decreased from 6.1 months [5.4–NA] when baseline was the reference score to 3.5 [3.2–6.0] when the reference was the best previous score (Fig. 3a).

Among the 375 patients included with at least one breast symptoms score, 228 and 284 patients presented breast symptom deterioration as compared to the baseline score and the best previous score, respectively. The median TTD increased from 0.2 months [0.2–2.8] when recalibration was not taken into account to 2.8 [2.8–3.0] when it was taken into account (Fig. 3b).

Among the 375 patients included with at least one arm symptoms score, 214 and 247 patients presented arm symptoms deterioration as compared to the baseline score and to the best previous score, respectively. The median TTD increased from 2.9 months [0.4–3.1] when recalibration was not taken into account to 6.0 [3.6–6.0] when it was.

Table 2 Results of the Kaplan–Meier estimation of the time to deterioration (TTD) for each QLQ-C30 score and QLQ-BR23 score with the baseline score or the best previous score as the reference score regarding breast cancer study (study #1)

	TTD baseline score		TTD best previous score	
	n (events)	Median in months (95 % CI)	*n* (events)	Median in months (95 % CI)
QLQ-C30				
Global health status	376 (224)	3.0 (2.8–3.0)	376 (263)	3.0 (2.9–3.0)
Physical functioning	376 (255)	0.2 (0.2–2.8)	376 (290)	0.4 (0.2–2.9)
Role functioning	375 (235)	3.0 (3.0–3.0)	375 (262)	3.0 (3.0–3.0)
Emotional functioning	377 (153)	6.1 (6.0–NA)	377 (232)	5.6 (3.2–5.9)
Social functioning	377 (193)	3.1 (3.0–5.9)	377 (221)	3.1 (3.0–5.4)
Cognitive functioning	377 (160)	6.1 (5.4–NA)	377 (197)	3.5 (3.2–6.0)
Fatigue	374 (248)	2.7 (0.2–3.0)	374 (282)	2.9 (0.4–3.0)
Pain	377 (234)	3.0 (0.6–3.0)	377 (268)	4.0 (2.8–3.0)
Nausea and vomiting	375 (123)	7.0 (6.1–NA)	375 (139)	7.0 (6.1–NA)
Dyspnea	375 (126)	6.2 (6.1–NA)	375 (164)	6.1 (6.0–6.2)
Insomnia	374 (141)	6.1 (6.0–NA)	374 (194)	6.0 (5.7–6.0)
Appetite loss	375 (106)	NA (6.3–NA)	375 (124)	6.5 (6.3–NA)
Constipation	377 (147)	6.2 (6.0–NA)	377 (173)	6.0 (5.9–6.4)
Diarrhea	375 (59)	NA (6.5–NA)	375 (81)	6.5 (6.4–NA)
Financial difficulties	376 (70)	NA (6.4–NA)	376 (78)	NA (6.4–NA)
QLQ-BR23				
Body image	376 (207)	3.0 (3.0–3.1)	376 (236)	3.0 (3.0–3.2)
Sexual functioning	354 (71)	6.4 (6.3–NA)	354 (118)	6.2 (6.1–6.4)
Sexual enjoyment	224 (21)	7.4 (6.4–NA)	224 (45)	6.4 (6.2–NA)
Future perspective	375 (90)	7.0 (6.6–NA)	375 (165)	6.1 (6.0–6.1)
Systemic therapy side effects	376 (194)	3.1 (3.0–3.4)	375 (233)	3.1 (3.0–3.2)
Breast symptoms	375 (228)	0.2 (0.2–2.8)	375 (284)	3.0 (2.8–3.0)
Arm symptoms	375 (214)	2.9 (0.4–3.1)	375 (247)	6.0 (3.6–6.0)
Upset by hair loss	194 (16)	3.3 (3.1–NA)	194 (38)	6.3 (6.2–NA)

Pancreatic cancer

Between October 2007 and May 2011, 98 patients were enrolled in 10 French centers. Mean age was 62 years (SD = 8.4). The baseline characteristics of the patients are summarized in supplementary Table B.

At baseline, 34 patients (69.4 %) completed the QLQ-C30 questionnaire in Arm 1 (gemcitabine + FOLFIRI 3) and 30 patients (61.2 %) in Arm 2 (gemcitabine alone) (supplementary Table C).

The TUDD as compared to the baseline score with a 5-point MCID or death was retained for the primary analysis. The Kaplan–Meier curves showing TUDD for the physical functioning and pain scales are shown in Fig. 4.

Patients in Arm 1 (gemcitabine + FOLFIRI 3) seem to present a longer TUDD than those in Arm 2 (gemcitabine alone) for each HRQoL score (Table 3).

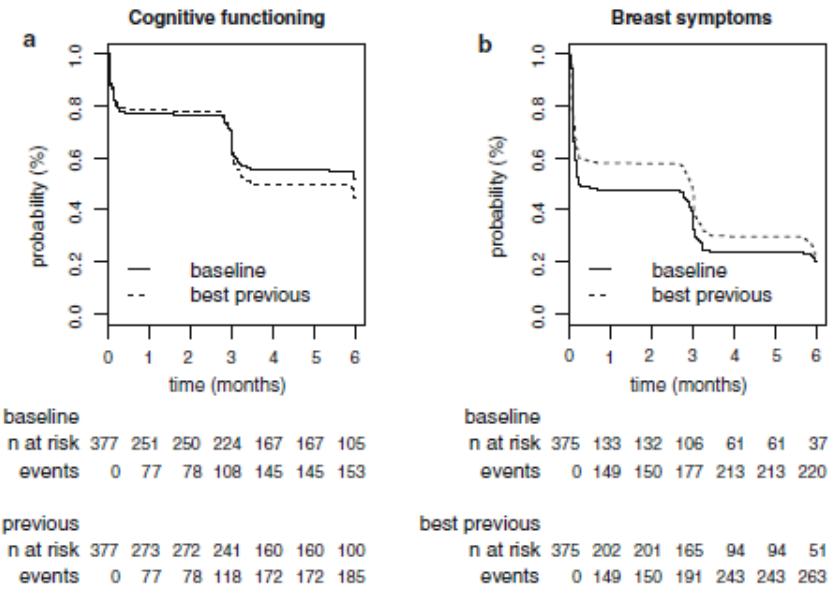
Whatever the definition applied, patients in Arm 1 (gemcitabine + FOLFIRI 3) presented a longer TUDD of insomnia than those of Arm 2 (gemcitabine alone) with HR < 1.

Regarding TUDD definitions integrating death or not, patients in Arm 1 (gemcitabine + FOLFIRI 3) presented a longer TUDD than those in Arm 2 (gemcitabine alone) for physical functioning, but this trend was no longer significant when we considered patients with no follow-up as having deteriorated at baseline.

Discussion

Definitions of deterioration applied in this paper, such as TTD compared to baseline score in breast cancer, and TUDD according to the definition of Bonnetain et al. in the pancreatic cancer study, have also been applied in other studies [39, 40]. This demonstrates the didactic nature of this approach.

Different definitions of TTD have been proposed and investigated in this paper. According to the definition applied, results can change and this precludes



baseline							
n at risk	377	251	250	224	167	167	105
events	0	77	78	108	145	145	153

best previous							
n at risk	377	273	272	241	160	160	100
events	0	77	78	118	172	172	185

baseline							
n at risk	375	133	132	106	61	61	37
events	0	149	150	177	213	213	220

best previous							
n at risk	375	202	201	165	94	94	51
events	0	149	150	191	243	243	263

Fig. 3 Time to HRQoL score deterioration curves with a 5-point MCID for breast cancer (study #1) with baseline score or best previous score as the reference score for cognitive functioning (CF) (*panel A*), and breast symptoms (BS) (*panel B*)

comparison of results between oncology clinical trials. The multiplicity of possible event definitions is a limitation of TTD analysis, as it can change the conclusions drawn from the same study. For this reason, it is essential to achieve a consensus. Moreover, if interval estimation of survival analysis is used, the "real" deterioration time is unknown, and as a result, the TTD will be overestimated, but biological markers such as progression-free survival also use this estimation method. An alternative is under investigation, for example, with patients completing the HRQoL questionnaire when they perceive a change.

In this paper, we report the results of TTD analyses according to different therapeutic situations (adjuvant or metastatic) and cancer sites (breast and pancreatic cancers). The impact of some challenges of longitudinal HRQoL analysis on TTD is also studied, namely occurrence of RS in breast cancer study and missing data in pancreatic cancer. We adjusted the definition of deterioration and the choice of the reference score according to the problem being studied.

In the breast cancer study, we noted that the choice of the reference score impacted on the median TTD. When the best previous score was used as the reference, rather than the baseline score, the median TTD of cognitive functioning decreased while that of the breast and arm symptoms increased. The median TTD is sensitive to the choice of reference score. One limitation of this study is the number of HRQoL assessments. Only four assessments of HRQoL during the study were planned. In the pancreatic cancer study, results were slightly different according to the definition applied. Regarding TUDD definitions integrating death or not, patients in Arm 1 (gemcitabine + FOLFIRI 3) presented a significantly longer TUDD than those of Arm 2 (gemcitabine alone) for physical functioning, but this trend was no longer significant when we considered patients with no follow-up as deteriorated at baseline.

In early breast cancer (study #1), the TTD definition applied, using the best previous score as the reference, has the advantage of taking into account the occurrence of the recalibration component of RS. The occurrence of short-term recalibration in this study was previously demonstrated [33]; thus, we had to adjust the method of longitudinal analysis according to the change in the patients' internal standards. Different methods of assessing RS exist [41–43]. However, the challenge is to take into account the occurrence of the RS effect in longitudinal analysis in order to estimate the true change. The "then-test" method, which assesses patients' pretest HRQoL levels retrospectively, is the most popular method to assess RS [44]. However, this method is time-consuming, and given its retrospective nature, the then-test is susceptible to recall bias [45]. The TTD approach has the advantage of taking recalibration into account without additional questionnaires, by using changing scores as a reference. Currently, few longitudinal methods can integrate the occurrence of a RS effect. Structural equation modeling can separate true change from RS effect [42, 46]. However, due to the complexity of this method, it is difficult to propose a simple interpretation of these models to clinicians.

The TTD approach is suitable for different therapeutic situations. Indeed, using the pancreatic cancer study, we

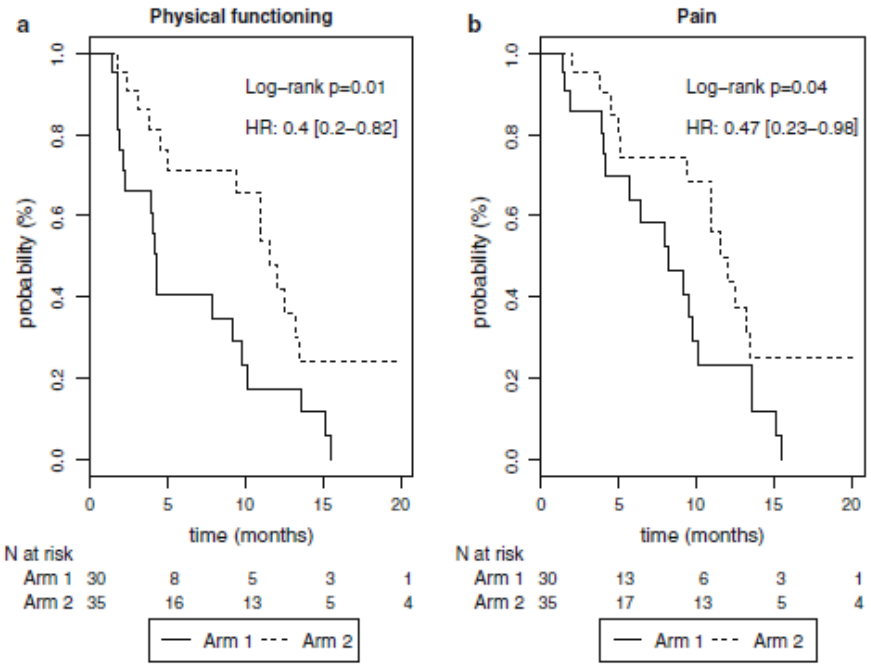


Fig. 4 Kaplan–Meier survival curves for the time until definitive deterioration or death for the pancreatic cancer study (study #2)

integrated the metastatic component as a definitive deterioration with death as an event.

Many definitions of deterioration have been proposed in this paper. The choice of the event definition is essential, because it may induce different results. However, there is currently no recommendation or consensus on this point. Consequently, TTD reflects heterogeneity. In the adjuvant setting, we thus recommend using the TTD; and in the advanced or metastatic setting, we recommend using the TUDD with or without death as an event. The baseline score could be considered as the reference score if there is no evidence of a RS effect. If a RS is likely to occur, we recommend using the best previous score or the previous score as the reference score in the TTD analysis.

As in other statistical methods for longitudinal analysis, the TTD approach can handle the occurrence of missing data by making some underlying assumptions, either by considering that the HRQoL level is constant for intermittent missing data, or by considering the missing HRQoL score as revealing the deterioration of the patient's health status. Few statistical methods handle missing data in longitudinal studies of HRQoL, and these methods are rarely applied due to their complexity. Pattern mixture models have been proposed to analyze longitudinal HRQoL with missing data [10, 25]. However, the number of patterns may be considerable and makes difficult the estimation of the model parameters for each plan. In this way, the TTD approach seems to be more appropriate than GLMM with pattern mixture for studies with many HRQoL assessments, although these two approaches measure different concepts, and thus, TTD cannot be a substitute for GLMM. In the pancreatic cancer study, we considered patients with no follow-up measure as having deteriorated since baseline. Further research is needed to take into account missing data profiles in TTD analyses. We are currently developing a method to use in conjunction with TTD to take into account missing not-at-random data using a method derived from a propensity score.

Results of TTD analysis could be more suitable than GLMM for clinicians, who are familiar with survival analysis, with HR, and log-rank test. However, both GLMM and TTD rely on the definition of MCID to be effective from a clinical point of view. Thus, these methods share the same limitation deriving from the lack of consensus around the MCID definition. Longitudinal results should have the ability to translate findings into information that decision-makers find understandable and compelling. At this time, despite available statistical approaches, results are poorly utilized to change standards of care, mainly due to the lack of standardization and the failure to propose clinical meaningful results.

An ongoing project aims to compare TTD and GLMM using a simulation study [47, 48]. The objective of this

Table 3 Results of the Kaplan–Meier estimation of the time until definitive deterioration (TUDD) for each QLQ-C30 score[1] and comparison between arms regarding pancreatic cancer study (study #2)

		TUDD baseline				TUDD baseline or death				TUDD baseline or death or no follow-up			
		n (events)	Median (CI 95 %)	Log-rank	HR (CI 95 %)	n (events)	Median (CI 95 %)	Log-rank	HR (CI 95 %)	n (events)	Median (CI 95 %)	Log-rank	HR (CI 95 %)
GHS	Gemcitabine alone	30 (6)	4.34 (2.2–NA)		1	30 (18)	7.92 (4.21–13.6)		1	30 (25)	4.27 (2.2–9.72)		1
	Gemcitabine + folfiri.3	33 (6)	3.22 (1.97–NA)	0.82	1.14 (0.37–3.54)	33 (16)	9.46 (3.81–NA)	0.45	0.77 (0.38–1.55)	33 (28)	3.22 (1.15–12.06)	0.95	1.02 (0.58–1.76)
PF	Gemcitabine alone	30 (9)	2.27 (1.91–NA)		1	30 (19)	4.27 (2.27–10.15)		1	30 (26)	3.98 (1.84–9.13)		1
	Gemcitabine + folfiri.3	35 (5)	12.06 (11.6–NA)	0.02	0.23 (0.06–0.85)	35 (17)	11.6 (9.46–26.25)	0.01	0.4 (0.2–0.82)	35 (30)	4.5 (0.03–12.06)	0.21	0.7 (0.41–1.22)
RF	Gemcitabine alone	30 (6)	4.27 (1.91–NA)		1	30 (18)	6.47 (4.04–13.57)		1	30 (25)	4.04 (1.84–9.72)		1
	Gemcitabine + folfiri.3	35 (5)	12.06 (7.36–NA)	0.17	0.39 (0.1–1.57)	35 (16)	11.01 (7.36–22.57)	0.15	0.6 (0.29–1.21)	35 (29)	4.5 (0.03–12.06)	0.70	0.9 (0.52–1.56)
EF	Gemcitabine alone	30 (9)	4.27 (2.2–NA)		1	30 (19)	5.75 (4.04–9.72)		1	30 (26)	4.21 (1.91–8.25)		1
	Gemcitabine + folfiri.3	33 (8)	5.06 (1.91–NA)	0.60	0.76 (0.28–2.07)	33 (18)	11.01 (3.81–22.57)	0.05	0.5 (0.25–1.02)	33 (29)	3.68 (0.92–12.48)	0.32	0.75 (0.43–1.32)
CF	Gemcitabine alone	29 (5)	4.34 (4.01–NA)		1	29 (17)	8.25 (4.27–13.57)		1	29 (23)	4.34 (4.01–9.72)		1
	Gemcitabine + folfiri.3	32 (7)	6.01 (3.81–NA)	0.95	0.97 (0.29–3.26)	32 (17)	9.46 (5.03–13.21)	0.69	0.87 (0.43–1.73)	32 (28)	3.81 (1.84–10.97)	0.70	1.12 (0.64–1.96)
SF	Gemcitabine alone	30 (7)	4.27 (2.2–NA)		1	30 (18)	7.92 (4.04–13.57)		1	30 (25)	4.04 (1.91–9.72)		1
	Gemcitabine + folfiri.3	33 (8)	3.22 (2.07–NA)	0.95	1.03 (0.36–2.95)	33 (19)	9.46 (3.22–13.47)	0.37	0.73 (0.37–1.45)	33 (30)	3.09 (1.15–11.01)	0.85	0.95 (0.55–1.64)
FA	Gemcitabine alone	30 (6)	4.34 (2.2–NA)		1	30 (18)	7.92 (4.21–13.57)		1	30 (25)	4.21 (1.91–9.72)		1
	Gemcitabine + folfiri.3	35 (6)	11.6 (8.57–NA)	0.15	0.38 (0.09–1.52)	35 (17)	11.01 (8.57–13.47)	0.13	0.59 (0.3–1.18)	35 (30)	4.5 (0.03–11.6)	0.67	0.89 (0.51–1.53)
NV	Gemcitabine alone	30 (5)	NA (2.33–NA)		1	30 (18)	8.25 (4.04–13.57)		1	30 (25)	6.47 (2.33–9.49)		1
	Gemcitabine + folfiri.3	33 (3)	NA (NA–NA)	0.40	0.54 (0.13–2.29)	33 (14)	12.48 (9.46–NA)	0.04	0.47 (0.22–0.99)	33 (25)	5.03 (1.81–13.21)	0.30	0.74 (0.42–1.31)
PA	Gemcitabine alone	30 (4)	5.75 (5.75–NA)		1	30 (18)	8.25 (5.75–13.57)		1	30 (25)	5.75 (3.98–9.72)		1
	Gemcitabine + folfiri.3	35 (4)	12.06 (11.6–NA)	0.16	0.31 (0.05–1.76)	35 (16)	11.6 (10.97–NA)	0.04	0.47 (0.23–0.98)	35 (29)	5.04 (0.03–12.06)	0.41	0.79 (0.45–1.38)
DY	Gemcitabine alone	30 (4)	4.86 (2.33–NA)		1	30 (17)	8.69 (4.86–13.6)		1	30 (24)	6.47 (2.33–9.72)		1
	Gemcitabine + folfiri.3	35 (3)	15.64 (NA–NA)	0.15	0.3 (0.05–1.69)	35 (15)	12.48 (9.46–NA)	0.02	0.41 (0.19–0.89)	35 (28)	5.03 (0.03–13.21)	0.38	0.76 (0.43–1.34)
In	Gemcitabine alone	30 (8)	4.37 (1.91–NA)		1	30 (19)	5.75 (3.98–9.2)		1	30 (26)	4.04 (1.84–8.25)		1
	Gemcitabine + folfiri.3	35 (2)	12.06 (12.06–NA)	0.01	0.07 (0.01–0.56)	35 (13)	12.06 (10.97–NA)	<0.01	0.24 (0.11–0.54)	35 (26)	5.03 (0.03–13.21)	0.04	0.56 (0.31–1)
AP	Gemcitabine alone	30 (5)	4.83 (4.27–NA)		1	30 (19)	7.92 (4.27–13.57)		1	30 (26)	4.83 (3.98–9.49)		1
	Gemcitabine + folfiri.3	35 (4)	12.06 (12.06–NA)	0.14	0.31 (0.06–1.62)	35 (15)	12.06 (10.97–NA)	0.04	0.47 (0.23–0.98)	35 (28)	5.03 (0.03–12.48)	0.41	0.79 (0.45–1.38)

[1] The QLQ-C30 measures five functional scales (*PF* physical functioning, *RF* role functioning, *EF* emotional functioning, *CF* cognitive functioning, *SF* social functioning), *GHS* global health status, and nine symptom scales (*FA* fatigue, *NV* nausea and vomiting, *PA* pain, *DY* dyspnea, *In* insomnia, *Ap* appetite loss, constipation, diarrhea, and financial difficulties). Results for constipation, diarrhea, and financial difficulties are not shown

project is to propose a standard for longitudinal HRQoL analysis in oncology according to therapeutic situations and cancer sites.

To reach the goal of standardized longitudinal analysis methods for HRQoL, we purport that RECIST criteria for HRQoL regarding TTD are required. We propose the first components of the RECIST criteria here: (1) TTD and TUDD in the adjuvant and advanced/metastatic settings, respectively, with baseline score as a reference, and (2) with the best previous score or the previous score as a reference if RS effect is likely to occur. Further work is needed to achieve a consensus for each cancer setting and tumor site. Moreover, additional investigations are still required regarding the MCID determination to achieve consensus on a definition for MCID.

The TTD approach is already implemented in R software (submitted soon) to allow wider dissemination of these approaches and help move toward the goal of standardization.

At this time, the international ARCAD group ("Aide et Recherche en Cancérologie Digestive") supports the idea of developing RECIST criteria for HRQoL in colorectal cancer with liver metastasis and pancreatic cancer. Subsequently, HRQoL could then be considered as a co-primary endpoint along with a tumor parameters such as progression-free survival [49]. Future research is warranted on this subject [50]. For example, calculating the number of subjects required for a study with co-primary endpoints is still ongoing.

Conclusion

The TTD is a didactic and promising approach that we recommend for the longitudinal analysis of HRQoL in oncology, especially because of its capacity to handle RS and to provide results in a format that is familiar to clinicians.

Acknowledgments We thank Fiona Ecarnot for correcting the manuscript.

References

1. Beitz, J., Gnecco, C., & Justice, R. (1996). Quality-of-life end points in cancer clinical trials: The US food and drug administration perspective. *Journal of the National Cancer Institute Monographs*, (20), 7–9.
2. Johnson, J. R., & Temple, R. (1985). Food and drug administration requirements for approval of new anticancer drugs. *Cancer Treatment Reports, 69*(10), 1155–1159.
3. Lipscomb, J., Donaldson, M. S., Arora, N. K., Brown, M. L., Clauser, S. B., Potosky, A. L., et al. (2004). Cancer outcomes research. *Journal of the National Cancer Institute Monographs*, (33), 178–197.
4. Fairclough, D. L., Peterson, H. F., & Chang, V. (1998). Why are missing quality of life data a problem in clinical trials of cancer therapy? *Statistics in Medicine, 17*(5–7), 667–677.
5. Ross, L., Thomsen, B. L., Boesen, E. H., & Johansen, C. (2004). In a randomized controlled trial, missing data led to biased results regarding anxiety. *Journal of Clinical Epidemiology, 57*(11), 1131–1137.
6. Curran, D., Bacchi, M., Schmitz, S. F., Molenberghs, G., & Sylvester, R. J. (1998). Identifying the types of missingness in quality of life data from clinical trials. *Statistics in Medicine, 17*(5–7), 739–756.
7. Van Steen, K., Curran, D., & Molenberghs, G. (2001). Sensitivity analysis of longitudinal binary quality of life data with drop-out: an example using the EORTC QLQ-C30. *Statistics in Medicine, 20*(24), 3901–3920.
8. Cole, B. F., Bonetti, M., Zaslavsky, A. M., & Gelber, R. D. (2005). A multistate Markov chain model for longitudinal, categorical quality-of-life data subject to non-ignorable missingness. *Statistics in Medicine, 24*(15), 2317–2334.
9. Fairclough, D. L., Peterson, H. F., Cella, D., & Bonomi, P. (1998). Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. *Statistics in Medicine, 17*(5–7), 781–796.
10. Pauler, D. K., McCoy, S., & Moinpour, C. (2003). Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine, 22*(5), 795–809.
11. Troxel, A. B., Fairclough, D. L., Curran, D., & Hahn, E. A. (1998). Statistical analysis of quality of life with missing data in cancer clinical trials. *Statistics in Medicine, 17*(5–7), 653–666.
12. Liao, K., Freres, D. R., & Troxel, A. B. (2012). A transition model for quality-of-life data with non-ignorable non-monotone missing data. *Statistics in Medicine, 31*(28), 3444–3466.
13. Ubel, P. A., Peeters, Y., & Smith, D. (2010). Abandoning the language of "response shift": A plea for conceptual clarity in distinguishing scale recalibration from true changes in quality of life. *Quality of Life Research, 19*(4), 465–471.
14. Wiklund, I. (2004). Assessment of patient-reported outcomes in clinical trials: The example of health-related quality of life. *Fundamental & Clinical Pharmacology, 18*(3), 351–363.
15. Bullinger, M. (2002). Assessing health related quality of life in medicine. An overview over concepts, methods and applications in international research. *Restorative Neurology and Neuroscience, 20*(3–4), 93–101.
16. Gibbons, F. X. (1999). Social comparison as a mediator of response shift. *Social Science and Medicine, 48*(11), 1517–1530.
17. Sprangers, M. A., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Social Science and Medicine, 48*(11), 1507–1515.
18. Pan, A. W., Chen, Y. L., Chung, L. I., Wang, J. D., Chen, T. J., & Hsiung, P. C. (2012). A longitudinal study of the predictors of quality of life in patients with major depressive disorder utilizing a linear mixed effect model. *Psychiatry Research, 198*(3), 412–419.
19. Hunger, M., Doring, A., & Holle, R. (2012). Longitudinal beta regression models for analyzing health-related quality of life scores over time. *BMC Medical Research Methodology, 12*, 144.
20. Penar-Zadarko, B., Binkowska-Bury, M., Wolan, M., Gawelko, J., & Urbanski, K. (2013). Longitudinal assessment of quality of life in ovarian cancer patients. *European Journal of Oncology Nursing, 17*(3), 381–385.
21. Mantegna, G., Petrillo, M., Fuoco, G., Venditti, L., Terzano, S., Anchora, L. P., et al. (2013). Long-term prospective longitudinal evaluation of emotional distress and quality of life in cervical

cancer patients who remained disease-free 2-years from diagnosis. *BMC Cancer, 13*, 127.
22. Rathod, S., Gupta, T., Ghosh-Laskar, S., Murthy, V., Budrukkar, A., & Agarwal, J. (2013). Quality-of-life (QOL) outcomes in patients with head and neck squamous cell carcinoma (HNSCC) treated with intensity-modulated radiation therapy (IMRT) compared to three-dimensional conformal radiotherapy (3D-CRT): Evidence from a prospective randomized study. *Oral Oncology, 49*(6), 634–642.
23. Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine, 16*(20), 2349–2380.
24. Fairclough, D. L. (2010). *Design and analysis of quality of life studies in clinical trials*. Boca Raton: CRC Press.
25. Little, R. J., & Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics, 52*(1), 98–111.
26. Hogan, J. W., & Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine, 16*(1–3), 239–257.
27. Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., & Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics, 3*(2), 245–265.
28. Glas, C. A., Geerlings, H., van de Laar, M. A., & Taal, E. (2009). Analysis of longitudinal randomized clinical trials using item response models. *Contemporary Clinical Trials, 30*(2), 158–170.
29. De Ayala, R. (2009). *The theory and practice of item response theory*. New York: Guilford Press.
30. Bonnetain, F., Dahan, L., Maillard, E., Ychou, M., Mitry, E., Hammel, P., et al. (2010). Time until definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. *European Journal of Cancer, 46*(15), 2753–2762.
31. Hamidou, Z., Dabakuyo, T. S., Mercier, M., Fraisse, J., Causeret, S., Tixier, H., et al. (2011). Time to deterioration in quality of life score as a modality of longitudinal analysis in patients with breast cancer. *Oncologist, 16*(10), 1458–1468.
32. Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research, 1*(4), 274.
33. Dabakuyo, T. S., Guillemin, F., Conroy, T., Velten, M., Jolly, D., Mercier, M., et al. (2013). Response shift effects on measuring post-operative quality of life among breast cancer patients: A multicenter cohort study. *Quality of Life Research, 22*(1), 1–11.
34. Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute, 85*(5), 365–376.
35. Sprangers, M. A., Groenvold, M., Arraras, J. I., Franklin, J., te Velde, A., Muller, M., et al. (1996). The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: First results from a three-country field study. *Journal of Clinical Oncology, 14*(10), 2756–2768.
36. Andrykowski, M. A., Donovan, K. A., & Jacobsen, P. B. (2009). Magnitude and correlates of response shift in fatigue ratings in women undergoing adjuvant therapy for breast cancer. *Journal of Pain and Symptom Management, 37*(3), 341–351.
37. Fayers, P. M., Aaronson, N. K., Bjordal, K., Groenvold, M., Curran, D., Bottomley, A. ObotEQoLG. EORTC QLQ-C30 Scoring Manual (3rd edition). Brussels: EORTC 2001 ed2001.
38. Team, R. D. C. R. A language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing. ISBN 3-900051-07-0, http://www.R-project.org/.
39. Gourgou-Bourgade, S., Bascoul-Mollevi, C., Desseigne, F., Ychou, M., Bouche, O., Guimbaud, R., et al. (2013). Impact of FOLFIRINOX compared with gemcitabine on quality of life in patients with metastatic pancreatic cancer: Results from the PRODIGE 4/ACCORD 11 randomized trial. *Journal of Clinical Oncology, 31*(1), 23–29.
40. Wimberger, P., Gilet, H., Gonschior, A. K., Heiss, M. M., Moehler, M., Oskay-Oezcelik, G., et al. (2012). Deterioration in quality of life (QoL) in patients with malignant ascites: Results from a phase II/III study comparing paracentesis plus catumaxomab with paracentesis alone. *Annals of Oncology, 23*(8), 1979–1985.
41. Korfage, I. J., de Koning, H. J., & Essink-Bot, M. L. (2007). Response shift due to diagnosis and primary treatment of localized prostate cancer: A then-test and a vignette study. *Quality of Life Research, 16*(10), 1627–1634.
42. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*(3), 587–598.
43. Schwartz, C. E., & Sprangers, M. A. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science and Medicine, 48*(11), 1531–1548.
44. Sprangers, M. A., Van Dam, F. S., Broersen, J., Lodder, L., Wever, L., Visser, M. R., et al. (1999). Revealing response shift in longitudinal research on fatigue–the use of the thentest approach. *Acta Oncologica, 38*(6), 709–718.
45. McPhail, S., & Haines, T. (2010). Response shift, recall bias and their effect on measuring change in health-related quality of life amongst older hospital patients. *Health and Quality of Life Outcomes, 8*, 65.
46. Oort, F. J., Visser, M. R., & Sprangers, M. A. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research, 14*(3), 599–609.
47. Blanchin, M., Hardouin, J. B., Le Neel, T., Kubis, G., Blanchard, C., Mirallie, E., et al. (2011). Comparison of CTT and Rasch-based approaches for the analysis of longitudinal patient reported outcomes. *Statistics in Medicine, 30*(8), 825–838.
48. Sebille, V., Hardouin, J. B., Le Neel, T., Kubis, G., Boyer, F., Guillemin, F., et al. (2010). Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients–a simulation study. *BMC Medical Research Methodology, 10*, 24.
49. Booth, C. M., & Eisenhauer, E. A. (2012). Progression-free survival: Meaningful or simply measurable? *Journal of Clinical Oncology, 30*(10), 1030–1033.
50. Bonnetain, F., Bosset, J. F., Gerard, J. P., Calais, G., Conroy, T., Mineur, L., et al. (2012). What is the clinical benefit of preoperative chemoradiotherapy with 5FU/leucovorin for T3-4 rectal cancer in a pooled analysis of EORTC 22921 and FFCD 9203 trials: Surrogacy in question? *European Journal of Cancer, 48*(12), 1781–1790.

2.2. Développement d'un package R pour l'analyse longitudinale de la QdV selon la méthode du temps jusqu'à détérioration

Résumé

Introduction

L'analyse longitudinale de la qualité de vie relative à la santé (QdV) reste encore complexe et non standardisée. Le temps jusqu'à détérioration (TJD) d'un score de QdV a ainsi été proposé comme modalité d'analyse longitudinale de la QdV en cancérologie (Bonnetain *et al*, 2010).
Dans cette optique, le package QoLR, dédié à l'analyse longitudinale de la QdV en cancérologie, a été créé sous le logiciel R (Team, 2010).

Ce package est disponible sur le site « Comprehensible R Archive Network » (CRAN) via le lien : **http://CRAN.R-project.org/package=QoLR.**

Méthodes

Le package QoLR permet de déterminer le TJD, définitive (état absorbant) ou non, d'un score de QdV. Différentes définitions de détérioration ont ainsi été investiguées en faisant varier :

- le choix du score de référence (score à l'inclusion, meilleur score antérieur, score immédiatement précédent),
- le choix de la différence minimale cliniquement importante (DMCI),
- et dépendant de la construction du score (la détérioration correspond à une augmentation ou à une diminution du score,
- intégrant ou non le décès comme évènement,
- prenant en compte ou non l'absence de mesure de suivi comme évènement.

D'autre part, le scoring des questionnaires de QdV de l'EORTC a également été implémenté selon la méthode d'imputation simple par la moyenne et selon les recommandations du manuel de scoring de l'EORTC (Fayers PM *et al*, 2001).

Résultats

Une fonction a été créée pour le scoring de chaque questionnaire et module de QdV de l'EORTC. Chacune de ces fonctions créée un data frame avec l'identifiant des patients et les scores de QdV estimés selon les recommandations du manuel de scoring de l'EORTC (Fayers

PM *et al*, 2001). Deux fonctions ont été créées pour l'évaluation du TJD : une fonction pour le TJD et une seconde pour le TJD définitif. Plusieurs scores de QdV peuvent être analysés dans une même procédure en précisant, pour chaque score, si la détérioration correspond à une augmentation ou à une diminution du score.

Des analyses de sensibilité sont proposées en option: en considérant le décès comme évènement, les patients sans mesure de suivi en détérioration dès l'inclusion et/ou en faisant varier la DMCI. Ces analyses peuvent également être réalisées en parallèle de l'analyse principale dans une seule et même procédure. Le résultat de ces fonctions est un « dataframe » avec l'identifiant du patient, une variable dichotomique ***event*** indiquant si le patient est détérioré ou non, et une variable ***time*** correspondant au temps jusqu'à détérioration ou censure. Un indicateur du score étudié et éventuellement de la DMCI et de l'analyse de sensibilité correspondante sont ajoutés à chaque variable ***time*** et ***event***.

QoLR contient également une interface graphique permettant d'obtenir les courbes de TJD selon la méthode de Kaplan-Meier, par bras de traitement, avec en option l'affichage du nombre de patients à risque et du nombre d'évènements cumulés à intervalle de temps réguliers. Une fonction du package permet également de réaliser l'analyse de TJD et d'exporter les résultats par bras de traitement sous la forme d'un tableau Excel avec le nombre d'évènements, la médiane de détérioration par bras de traitement ainsi que le Hazard Ratio et la P-value du test du Log-Rank. Deux jeux de données fictifs ont également été ajoutés au package afin de tester les différents programmes. Un fichier d'aide est fourni avec le package pour guider l'utilisateur dans le choix des fonctions et des paramètres (Annexe C).

Conclusion

QoLR est le premier package R dédié à l'analyse de la QdV. Son implémentation permet de réaliser simplement des analyses de QdV longitudinales selon la méthode du TJD et va permettre une large utilisation et diffusion de cette approche.

<u>Article:</u> QoLR: An R Package for the Longitudinal Analysis of Health-Related Quality of Life

Article en révision dans *Journal of Statistical Software*

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

QoLR: An R Package for the Longitudinal Analysis of Health-Related Quality of Life in Oncology

Amélie Anota
University Hospital of Besançon

Marion Savina
Bergonie Institute of Bordeaux

Caroline Bascoul-Mollevi
Cancer Institute of Montpellier

Franck Bonnetain
University Hospital of Besançon

Abstract

Health-related quality of life has an increasing importance in clinical trials over the past two decades. The R package **QoLR** is a recently developed package for the longitudinal analysis of health-related quality of life in oncology. This package contains the scoring of most of the European Organisation for Research and Treatment of Cancer quality of life questionnaires and some programs to analyze the time to a health-related quality of life score deterioration as a modality of longitudinal analysis in oncology.

Keywords: health-related quality of life, longitudinal analysis, time to deterioration, R package **QoLR**.

1. Introduction

1.1. Health-related quality of life

Health-related quality of life (HRQOL) is a subjective clinical endpoint that has had an increasing importance in clinical trials in oncology over the past two decades (Osoba 2011). Although overall survival is still considered as the primary objective and the primary endpoint in many studies, most clinical trials now integrate HRQOL as an endpoint in order to investigate the clinical benefit for the patient, it even seems that HRQOL could become the primary endpoint in metastatic settings.

1.2. Assessment

The HRQOL is mainly studied through validated questionnaires filling out by the patients at several times during the study. In oncology, the European Organisation for Research and Treatment of Cancer (EORTC) has developed a core questionnaire QLQ-C30 to assess HRQOL level of cancer patients with 30 questions (items) and some supplementary modules for disease specific treatment measurements (Aaronson, Ahmedzai, Bergman, Bullinger, Cull, Duez, Filiberti, Flechtner, Fleishman, de Haes *et al.* 1993).

The EORTC HRQOL questionnaires are composed of a set of items with several categories of responses, usually constructed on a Likert scale (e.g.,: "Not at all"/"a little"/"quite a bit"/"very much", coded 1/2/3/4). These questions allow the estimation of one or more HRQOL dimensions through the calculation of scores. Generally, a score is the weighted sum of patients' items responses regarding one dimension and can be calculated if at least half of the items are answered. The scores are often normalized on a 0-100 scale in order to facilitate the comparison with one or more supplementary HRQOL questionnaires.

For the most of the HRQOL questionnaires, a high score corresponds to a high level of functioning for a functional scale and to a high presence of symptom for a symptomatic scale. Conversely, a low score corresponds to a low level of functioning for a functional scale and to a low presence of symptom for a symptomatic scale. In this way, patients with a high level of HRQOL should have high scores for functional scales and low scores for symptomatic scales. For questionnaires which included an item about global health or global HRQOL (as for the EORTC QLQ-C30), the score obtained is coded like a functional scale.

The method to calculate the scores of the EORTC HRQOL questionnaire is defined in the EORTC scoring manual (Fayers, Aaronson, Bjordal, Curran, and Grønvold 1999). Briefly, all the scores of the EORTC HRQOL questionnaires are obtained with the same procedure:

Let $I_1, ..., I_n$ be the n items answers to the studied dimension.

The first step, common for functional and symptomatic scales, is the calculation of a raw score (RS), i.e., the mean of the items:

$$RS = \frac{I_1 + ... + I_n}{n} \tag{1}$$

If there are some missing items, the denominator equals to the number of non-missing items. If more than half of the items are missing, the raw score cannot be calculated and then the dimension score is missing.

The second step is a linear transformation of the scores to a 0 - 100 scale to obtain a score S:

- for functional scales:
$$S = (1 - \frac{RS - 1}{range}) \times 100 \tag{2}$$
- for symptomatic scales:
$$S = (\frac{RS - 1}{range}) \times 100 \tag{3}$$
- for global health status:
$$S = (1 - \frac{RS - 1}{range}) \times 100 \tag{4}$$

1.3. Longitudinal analysis

The longitudinal analysis of HRQOL is a major challenge due to different parameters that we have to take into account in the analysis.

Firstly, the longitudinal assessment of HRQOL may be compromised by the presence of missing data. Two types of missing data can occur: intermittent missing data (e.g., if a patient forgot to complete one or more questions at one measurement time) or monotone missing data (e.g., if a patient dropped out before the end of the study). The presence of missing data can be informative or non-informative of patient's health status. For example, if a patient dropped out the study due to a disease progression, monotone missing data occur and can provide information for patient's HRQoL level: the patient HRQOL level is likely to have decreased. Thus, missing data are missing not at random. Regarding intermittent missing data, in some circumstances, we can suppose that the patient's HRQOL level remains unchanged between two available measures (missing at random or completely at random).

Secondly, the self-assessment of HRQOL is subjective, i.e., it is dependent on the patient's internal standards and definition of HRQOL (Wiklund 2004; Bullinger 2002; Ubel, Peeters, and Smith 2010). However, patients' accommodation to treatments toxicity and their acceptance of the disease may induce that they do not necessary assess their HRQOL with the same criteria at all measurement times. These changes can be reflected by a response shift effect (Gibbons 1999; Sprangers and Schwartz 1999; Korfage, de Koning, and Essink-Bot 2007) and can bias the interpretation of longitudinal HRQOL analysis if it is not taken into account in the study design or at least in the analysis' method (Ahmed, Mayo, Corbiere, Wood-Dauphinee, Hanley, and Cohen 2005).

Response shift is defined as "a change in the meaning of one's self-evaluation of a target construct as a result of: (a) a change in the respondent's internal standards of measurement (i.e., scale recalibration); (b) a change in the respondent's values i.e., the importance of component domains constituting the target construct) or (c) a redefinition of the target construct (i.e., reconceptualization)" (Sprangers and Schwartz 1999). In this way, the choice of the level of HRQOL reference to qualify a change like deterioration can be a major concern: the baseline score is not systematically the reference score.

Finally, the longitudinal analysis of HRQOL involves the question of the minimal clinically important difference (MCID). Indeed, there is a difference between the notion of "statistically significant" and "clinically significant". Osoba *et al.* have demonstrated that a MCID of 5 points is a small change, a change between 10 and 20 points is moderated and more than 20 points is a high change (Osoba, Rodrigues, Myles, Zee, and Pater 1998).

Due to the complexity of the longitudinal assessment of HRQOL, there is still no gold standard to analyze HRQOL over time in oncology. Moreover, another challenge of statistical methods to analyze longitudinal HRQOL is to propose some meaningful results for the clinicians. There is a need to develop statistical methods adapted for the decision makers. Longitudinal results should have the ability to translate findings into information that decision makers find understandable and compelling. Thus, different methods to analyze longitudinal HRQOL have been proposed (Pan, Chen, Chung, Wang, Chen, and Hsiung 2012; Hunger, Döring, and Holle 2012; Penar-Zadarko, Binkowska-Bury, Wolan, Gawelko, and Urbanski 2012; Cnaan, Laird, and Slasor 1997). The most widely used is the linear mixed model (Diggle 1988). Survival analysis approach like the time to deterioration in a HRQOL score has recently been proposed as a modality of longitudinal HRQOL analysis in cancer patients (Bonnetain, Dahan, Maillard,

Ychou, Mitry, Hammel, Legoux, Rougier, Bedenne, and Seitz 2010; Hamidou, Dabakuyo, Mercier, Fraisse, Causeret, Tixier, Padeano, Loustalot, Cuisenier, Sauzedde *et al.* 2011).

The linear mixed model is optimal for a study design with 2 to 5 measurement times (Fairclough 2010). In this model, the time is considered as a categorical variable. Moreover, these models are only adapted for studies whose HRQOL assessments are performed in some periods with few amplitude within patients. These models can take into account the missing data profiles by applying a pattern mixture model (Pauler, McCoy, and Moinpour 2003). However, these sub models are rarely applied mainly because of the complexity of the construction of the patterns. Moreover, at this time, these models do not deal with the occurrence of a response shift effect. Finally, these models can not provide results easy to understand for the clinician who is not familiar with the beta change and the mixed models.

Contrary to the linear mixed model, the time to deterioration (TTD) models can propose clinically meaningful results with hazard ratio and log-rank test (Bonnetain *et al.* 2010; Hamidou *et al.* 2011). Moreover, these models can handle the presence of missing data: when some intermittent missing data occur for one patient, we can consider that the patient's HRQOL level remains unchanged since the previous available measure; when we have monotone missing data for one patient, we can consider that the presence of the missing data is due to a deterioration of the patient's health status or not necessarily according to the therapeutic setting: adjuvant or advanced/metastatic. Finally, these models can take into account the occurrence of the recalibration component of the response shift effect by choosing different scores as the reference score to qualify the deterioration (Anota, Hamidou, Paget-Bailly, Chibaudel, Bascoul-Mollevi, Auquier, Westeel, Fiteni, Borg, and Bonnetain 2013).

The aim of this paper is to present the R (R Core Team 2014) package **QoLR** which is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=QoLR`. This package allows to calculate the scores of the EORTC QLQ-C30 and most of its modules and to determine the time to deterioration in a HRQOL score as a modality of longitudinal analysis.

2. The time to deterioration definitions

At this time, several definitions of TTD in a HRQOL score have been proposed depending on event definition and censoring rules. The event definition could be defined according to the reference score, the MCID, missing scores, including death or not. The most intuitive definition is the time from inclusion in the study to a first deterioration of the score with a MCID of k points at least as compared to baseline score (Hamidou *et al.* 2011). Patients with no deterioration before their drop-out of the study are censored at the time of the last HRQOL questionnaire completion or last follow-up. According to the construction of the score, the deterioration corresponds to an increase (e.g., for symptomatic scales of the EORTC questionnaires) or a decrease (e.g., for functional scales of the EORTC questionnaires) of the score. Between two available HRQOL scores, the level of HRQOL is supposed to be constant.

The notion of "deterioration" requires a reference score. Generally, the reference score is the baseline score (before randomization or at the inclusion in the absence of randomization). However, in order to take into account the occurrence of a response shift effect, the reference can also be

- the best level of HRQOL already experienced by the patient (i.e., the best previous

score);

- or the previous score HRQOL score for the patient (i.e., immediately preceding score).

The value of these scores can change over time according to the patient experience of treatment and disease evolution. Using these changing references instead of baseline measurement could be considered as an alternative to take into account the occurrence of the recalibration component of the response shift effect.

Furthermore we can consider for event definition, deterioration as definitive (i.e., absorbing state) or not depending on therapeutic setting. This induces two concepts: the TTD and the time until definitive deterioration (TUDD). Several definitions of TUDD have been proposed according to the notion of "definitive deterioration". The TUDD has been defined as:

1. the time from inclusion in the study to:
   - a first deterioration of k points at least as compared to the reference score,
   - with no further improvement of k points at least as compared to the reference score,
   - or if the patient dropped out (i.e., no more HRQOL data available) just after the deterioration was observed resulting in missing data (Bonnetain *et al.* 2010).
2. the time from inclusion in the study to:
   - a first deterioration of k points at least as compared to the reference score,
   - with maintaining this deterioration of k points at least for all following scores, i.e., the deterioration is observed for all the scores following the first deterioration,
   - or if the patient dropped out just after the deterioration was observed resulting in missing data.
3. the time from inclusion in the study to:
   - a first deterioration of k points at least as compared to the reference score,
   - with no further improvement of k points at least as compared to the score qualifying the deterioration (i.e., the score at the time of the first deterioration observed),
   - or if the patient dropped out just after the deterioration was observed resulting in missing data.

To illustrate, the following equations correspond to the three definitions of TUDD of a X score with a k-point MCID observed at time i as compared to the reference score X_{ref} assuming that X represents a functional scale (i.e., a deterioration is observed when the score decreases): $\exists i > 1; X_i \leq X_{ref} - k$ and:

1. $\forall j > i, X_j \leq X_{ref} + k$ for the first definition of TUDD;
2. $\forall j > i, X_j \leq X_{ref} - k$ for the second definition of TUDD;
3. $\forall j > i, X_j \leq X_i + k$ for the last definition of TUDD.

DEFINITION MCID= k points	REFERENCE SCORE X_{ref} at T_i				EVENT DEFINITION at T_i	
	Baseline score	Best previous score		Previous score	EVENT = decrease	EVENT = increase
		EVENT decrease	EVENT increase			
TTD $\geq$ k points	X_1	$\max_{j\leq i}(X_j)$	$\min_{j\leq i}(X_j)$	X_{i-1}	$T_i = \min_{j>1}(T_j)$, $X_j \leq X_{ref} - k$	$T_i = \min_{j>1}(T_j)$, $X_j \geq X_{ref} + k$
TUDD $\geq$ k points with no further improvement $\geq$ k points as compared to X_{ref}	X_1	$\max_{j\leq i}(X_j)$	$\min_{j\leq i}(X_j)$	X_{i-1}	$T_i = \min_{j>1}(T_j)$/ •$X_j \leq X_{ref} - k$ •$\forall m > i, X_m \leq X_{ref}+k$	$T_i = \min_{j>1}(T_j)$/ •$X_j \geq X_{ref} - k$ •$\forall m > i, X_m \geq X_{ref}-k$
TUDD $\geq$ k points with a deterioration observed at all times following time T_i	X_1	$\max_{j\leq i}(X_j)$	$\min_{j\leq i}(X_j)$	X_{i-1}	$T_i = \min_{j>1}(T_j)$/ •$X_j \leq X_{ref} - k$ •$\forall m > i, X_m \leq X_{ref}-k$	$T_i = \min_{j>1}(T_j)$/ •$X_j \geq X_{ref} - k$ •$\forall m > i, X_m \geq X_{ref}+k$
TUDD $\geq$ k points with no further improvement $\geq$ k points as compared to X_i	X_1	$\max_{j\leq i}(X_j)$	$\min_{j\leq i}(X_j)$	X_{i-1}	$T_i = \min_{j>1}(T_j)$/ •$X_j \leq X_{ref} - k$ •$\forall m > i, X_m \leq X_{ref}+k$	$T_i = \min_{j>1}(T_j)$/ •$X_j \geq X_i - k$ •$\forall m > i, X_m \geq X_i - k$

Table 1: Event definitions at time T_i according to the definition of time to score X deterioration as compared to the reference score X_{ref}

Table 1 indicates the event definition according to the retained definition of TTD or TUDD. The time until definitive HRQOL score deterioration is mainly applied in advanced or palliative setting. All-cause death can also be considered as an event if the patient did not experience deterioration before death. In this way, TUDD or death could be redefined as "HRQOL deterioration-free survival".

Patients with none score available are excluded from the time to deterioration analyses. Patients with no baseline score are usually censored at baseline and those with no follow up scores but with a baseline score are censored one day after baseline. As for other analyses of HRQOL, in the TTD analyses, some sensitivity analyses could be performed:

- Considering patients with no baseline score as events;
- Considering patients with no follow-up score as events;
- Varying the MCID.

As example, we can choose a MCID of 10 points instead 5 points initially fixed. Regarding the TUDD, if a patient experienced a definitive deterioration with a 10-point MCID but not a definitive deterioration with a 5-point MCID for one of the proposed definition, thus this

patient also presented a definitive deterioration with a 5-point MCID, and the event time will be the time of the 10-point deterioration. Indeed, for patients experiencing both a TUDD with a 5-point MCID and a TUDD with a 10-point MCID, the TUDD with a 5-point MCID must be the time of the first deterioration observed (5-point or 10-point MCID).

The following Figure 1 summarizes the different proposed definitions of TTD and TUDD. To have a valid observation, we need a date of HRQOL measure and a HRQOL score. When the reference score is the best previous score or the previous score, patients with no baseline score but with a post baseline score are kept in the TTD analyses if they have at least one follow up score available after the reference score. They are not censured at baseline. The first reference score is the first score available.

As TTD analyses belong to survival analyses, the TTD estimation could be calculated using the Kaplan-Meier or actuarial method and described using median and 95% confidence interval (CI). The Kaplan-Meier method is based on the intuitive idea that to be alive at time T, it requires to be alive just before time T and to do not die at time T (Goel, Khanna, and Kishore 2010). Contrary to Kaplan-Meier method, in actuarial method probabilities are estimated for fixed interval time, not determined by the date of observed death. Both methods can handle presence of censored data, i.e., patients still alive at the end of the study.

In time to deterioration analyses, the event is "the HRQoL score deterioration".

The Kaplan-Meier estimation is given by the next formula:

$$S(t) = \prod_{t_i \leq t} \frac{n_i - m_i}{n_i} \tag{5}$$

where $n_i = n_{i-1} - m_{i-1} - c_{i-1}$ and:

- n_i is the number of subject at risk at time i, i.e., the number of patients still in the study and who do not present a deterioration until time $i-1$;
- m_i is the number of event observed at time i, i.e., the number of patients experiencing a HRQoL score deterioration at time i;
- c_i is the number of censored patients at time i, i.e., the number of patients who drop out at time i and who did not experienced a HRQOL deterioration before.

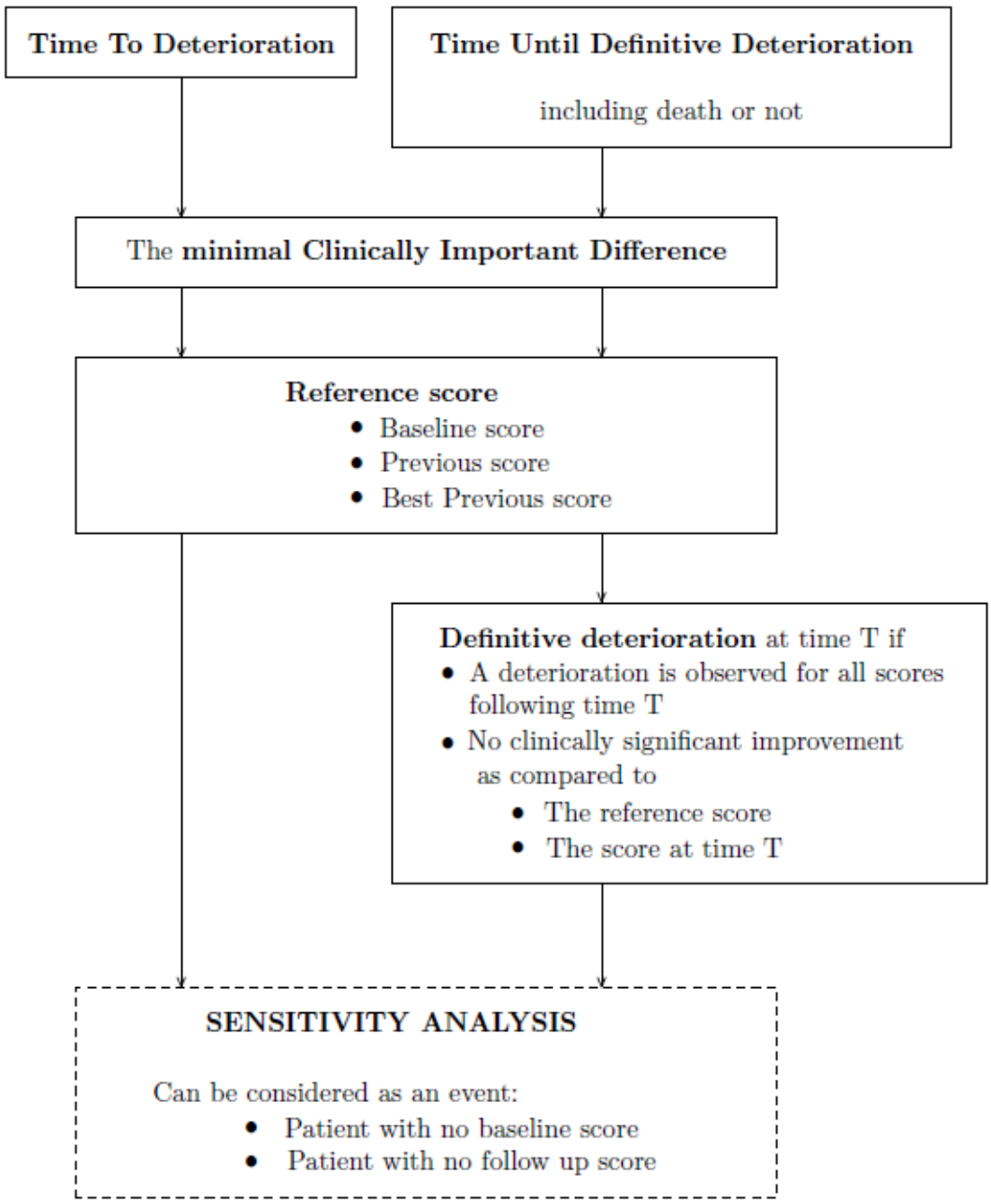


Figure 1: Flowchart of the choice of the definition of the time to a health-related quality of life score deterioration

TTD can then be compared according to treatment arm using the log-rank test and univariate Cox analyses to produce hazard ratio with 95% CI. Multivariate Cox regression can be applied to identify independent factors associated with TTD.

The log-rank test is a nonparametric test to compare the survival distributions of two samples A and B. The null hypothesis H_0 is the equality of the survival distribution in both groups A and B, i.e., the expected probability of the event at time i is the same in both groups. Under this null hypothesis, the theoretical probability of the event at time i is:

$$\mathsf{P}_i = (E_{Ai} + E_{Bi})/(V_{Ai} + V_{Bi}) \tag{6}$$

Where:

- E_{Ai} is the number of observed events in group A at time i;
- E_{Bi} is the number of observed events in group B at time i;
- V_{Ai} is the number of patients in group A which are not presenting the event at time i;
- V_{Bi} is the number of patients in group B which are not presenting the event at time i.

Then the log-rank statistic equals to:

$$\chi^2_{exp} = \frac{(\sum_i E_{Ai} - \sum_i(\mathsf{P}_i \times V_{Ai}))^2}{\sum_i(\mathsf{P}_i \times V_{Ai})} + \frac{(\sum_i E_{Bi} - \sum_i(\mathsf{P}_i \times V_{Bi}))^2}{\sum_i(\mathsf{P}_i \times V_{Bi})} \tag{7}$$

Under the null hypothesis, χ^2_{exp} is distributed according to a χ^2 distribution with one degree of freedom.

The Cox regression model link the instantaneously risk of event λ at time t with other covariates X_1, ..., X_n as follows:

$$\lambda(t, X_1, ..., X_n) = \lambda_0(t) \times \exp(\sum_{i=1}^{n} \beta_i X_i) \tag{8}$$

Where $\lambda_0(t)$ corresponds to a basis risk and corresponds to an instantaneously risk of event at time t when all covariates equals to 0.

3. Illustrations of the TTD and TUDD definitions

Event definition and censure rules depend on the definition of deterioration considered like it was illustrated in both Figure 1 and Table 1. Table 2 summarizes several situations for patients:

- the variable *id* is the patient's identification number;
- T_1 to T_5 correspond to five HRQOL assessments;
- a deterioration corresponds to a score decrease.

	HRQOL scores					TTD as compared to			TUDD as compared to								
									baseline			best previous			previous		
id	T_1	T_2	T_3	T_4	T_5	Baseline score	Best previous score	Previous score	1	2	3	1	2	3	1	2	3
1	61	59	55			T_3	T_3	T_3	T_3	T_3	T_3	T_3	T_3	T_3	T_3	T_3	T_3
2	61	61	56	51	59	T_3	T_3	T_3	T_3			T_3			T_3		
3	70	75	73	69	65	T_5	T_4		T_5	T_5	T_5	T_4	T_4	T_4			
4	51		58	50	55		T_4	T_4				T_4		T_4	T_4		T_4
5	64	61	58	47	59	T_3	T_3	T_4	T_3	T_3	T_3	T_3	T_3	T_3	T_4		
6	69	72		66			T_4	T_4				T_4	T_4	T_4	T_4	T_4	T_4
7	64	60	57	58	65	T_3	T_3		T_3			T_3					
8	55																
9	60	62	56	61	57		T_3	T_3				T_3	T_5	T_3	T_3		
10		61		53	61		T_4	T_4				T_4			T_4		

Table 2: HRQOL scores obtained at 5 times T_1 to T_5 for 10 patients and time events according to each definition of deterioration with a 5-point MCID

There are some missing scores for several measures, illustrating the issue of intermittent or monotone missing data in HRQOL studies.

Table 2 illustrates different scenarios according to the event definition. This table summarizes, for each definition of TTD and TUDD, the time of the events for patients experiencing a deterioration with a 5-point MCID. In other cases, patients are censored at the time of the last available HRQoL measure. For example, for the TTD as compared to the baseline score, patients ♯4 and ♯6 do not present a deterioration and are censored at T_5 and T_4 respectively. Patient ♯2 presents a deterioration as compared to the baseline score at T_3, but not definitive as compared to the score qualifying the deterioration (TUDD.3). Patient ♯9 presents a deterioration as compared to the best score at T_3: at T_3, the previous best score equals to 62 and the difference between this score and the score at T_3 equals to 6, greater than the MCID. However, the HRQoL level of the patient goes up to 61 and then the deterioration is not definitive as compared to the deteriorated score 56. At the last HRQOL assessment, HRQOL score equals to 57 and the previous best score is still 62. In this way, the deterioration observed at T_3 is definitive as compared to the deteriorated score. We recall that if a deterioration is followed by missing data (patient dropped out), the deterioration is definitive, whatever the definition of TUDD retained.

Two variables have then been created:

- a dummy variable event indicating if patient is deteriorated (event = 1) or not (event = 0);
- a time variable equals to the time between the date of baseline and the date of deterioration or censure.

The following Table 3 illustrates these two variables for the 10 patients of Table 2 and one definition of time to deterioration.

To illustrate, the patient ♯1 is in deterioration as compared to the baseline score (event = 1) and the time between the baseline date and the date of the deterioration equals to T_3 - T_1, then time = T_3 - T_1.

	TTD as compared to the baseline score		sensitivity analysis	
id	event	time	event	time
1	1	T_3 - T_1	1	T_3 - T_1
2	1	T_3 - T_1	1	T_3 - T_1
3	1	T_5- T_1	1	T_5- T_1
4	0	T_5- T_1	0	T_5- T_1
5	1	T_3 - T_1	1	T_3 - T_1
6	0	T_4- T_1	0	T_4- T_1
7	0	T_5- T_1	0	T_5- T_1
8	0	0	1	0
9	0	T_5- T_1	0	T_5- T_1
10	0	1	1	1

Table 3: Time to deterioration compared to the baseline score with a 5-point MCID for patients of Table 2 and sensitivity analysis considering patients with no baseline score or with no follow up score as events

Patient ♯8 has no baseline score:

- in the primary analysis, this patient is censured at baseline (event = 0);
- in the sensitivity analysis, this patient is in deterioration at baseline (event = 1).

In both cases, the time to deterioration equals to : time = T_1 - T_1 = 0.
In the same way, patient ♯10 has only a baseline score, no follow-up score:

- in the primary analysis, this patient is censured one day after baseline (event = 0);
- in the sensitivity analysis, this patient is in deterioration one day after baseline (event = 1).

In both cases, the time to deterioration equals to : time = $T_1 + 1$ - $T_1 = 1$.

4. Package QoLR

The **QoLR** package was developed to allow the longitudinal analysis of HRQOL according to the time to deterioration approach. The **QoLR** has dependencies with two R packages (**survival** (Therneau 2014) and **zoo** (Zeileis and Grothendieck 2005) packages).

4.1. Package structure

The **QoLR** package contains a set of functions to calculate the scores of the EORTC QLQ-C30 and most of its modules and two other programs to determine the time to deterioration in a HRQOL score as a modality of longitudinal analysis whatever the definition of deterioration retained. Other programs allow to print all of the results or in a csv file according to treatment arm and make all sensitivity analyses according to one reference score. A last program was created to obtain the TTD curves calculated according to the Kaplan-Meier estimation

method, displaying in option some information (number of patients at risk, cumulative number of events, hazard ratio and log-rank test if two treatment arms are compared).

For the convenience of the reader, we summarized all the main functions, with arguments and descriptions of our package **QoLR** in Table 4. In the following, we describe some functions of the software in detail.

Functions	Arguments	Description
`scoring.QLQC30`	`(data, id, items)`	Scoring of the EORTC QLQ-C30 questionnaire
`scoring.QLQBN20`	`(data, id, items)`	Scoring of the EORTC QLQ-BN20 module
`scoring.QLQBR23`	`(data, id, items)`	Scoring of the EORTC QLQ-BR23 module
`scoring.QLQCR29`	`(data, id, items)`	Scoring of the EORTC QLQ-CR29 module
`scoring.QLQCX24`	`(data, id, items)`	Scoring of the EORTC QLQ-CX24 module
`scoring.QLQEN24`	`(data, id, items)`	Scoring of the EORTC QLQ-EN24 module
`scoring.QLQHN35`	`(data, id, items)`	Scoring of the EORTC QLQ-H&N35 module
`scoring.QLQLC13`	`(data, id, items)`	Scoring of the EORTC QLQ-LC13 module
`scoring.QLQMY20`	`(data, id, items)`	Scoring of the EORTC QLQ-MY20 module
`scoring.QLQOES18`	`(data, id, items)`	Scoring of the EORTC QLQ-OES18 module
`scoring.QLQOG25`	`(data, id, items)`	Scoring of the EORTC QLQ-OG25 module
`scoring.QLQPR25`	`(data, id, items)`	Scoring of the EORTC QLQ-PR25 module
`scoring.QLQSTO22`	`(data, id, items)`	Scoring of the EORTC QLQ-STO22 module
`TTD`	`(X, score, ...)`	Estimation of the time to deterioration
`TUDD`	`(X, score, ...)`	Estimation of the time until definitive
`plotTTD`	`(time, event, ...)`	Kaplan-Meier curve of the TTD or TUDD
`write.TTD`	`(X, score, ...)`	Estimation of the TTD and print the results in a csv file
`write.TUUD`	`(X, score, ...)`	Estimation of the TUDD and print the results in a csv file

Table 4: Summary of the functions in the **QoLR** package

- *Function* `scoring.QLQC30()`

The first application of the **QoLR** package is the estimation of the scores of most of the EORTC HRQOL questionnaires like the QLQ-C30 cancer specific questionnaire.

The first argument of the function `scoring.QLQC30` and other functions for scoring is the name of the dataset with the items answers to the questionnaire (`X` parameter). The patient's identification number can be specified in `id` parameter. The last argument is the position of the items in the dataset (`items` parameter). By default, the items position are the first columns of the data frame. For example, the QLQ-C30 contains 30 items and `items = 1:30`.

The result is a data frame: each variable corresponds to a HRQOL score. The score names are the same than those used in the EORTC scoring manual (Fayers *et al.* 1999). If a patient's identification number was specified in the `id` parameter, then this variable was replicated in the data frame obtained.

- *Function* `TTD()`

This function is used to estimate the time to deterioration in a HRQOL score. To apply this function, the dataset must respect a general structure. The dataset `X` must be in long format

with the following variables in this order:

1. Patient's identification number;
2. Variable identifying the HRQOL assessment number;
3. Date of HRQOL measure;
4. HRQOL scores;
5. Other variables like the date of death or the treatment arm.

The dataset must also be sorted by patient's identification number and HRQOL measurement time. Dates must be in Julian format (i.e., number of days since a reference time point).

All definitions of TTD presented in this paper were programmed. Table 5 summarizes the arguments of this function and their possible values. According to the definition of deterioration retained, you must specify:

- The name of your dataset (`X`);
- The name of the HRQOL scores studied (`score = " "`);
- The value of the MCID (`MCID = 5`, as example);
- The reference score to qualify the deterioration which could be:
  - The baseline score (`ref.init = "baseline"`);
  - The best previous score (`ref.init = "best"`);
  - The previous score (`ref.init = "previous"`).
- If the deterioration corresponds to a decrease (`order = 1`) or an increase of the score (`order = 2`);
- If patients with no baseline score are censured (`no_baseline = "censure"`) or are in deterioration (`no_baseline = "event"`) since baseline;
- If patients with no follow-up score are censured (`no_follow = "censure"`) or are in deterioration (`no_follow = "event"`) one day after baseline;
- If death is considered as an event, you must specify the name of the variable in your dataset `X` which contains the date of death for patients who died during the study and with missing values for patients still alive at the end of the study.

An option (`sensitivity = TRUE`) allows to performed all sensitivity analyses available in one application of the `TTD()` function.

As example, if you fixed `no_baseline = "censure"`, `no_follow = "censure"` and `sensitivity = TRUE` then:

- A first analysis is conducted censuring patients with no baseline and those with no follow-up score;

- Sensitivity analysis ♯1: A sensitivity analysis is conducted considering these patients in deterioration.

If in addition to these parameters, the variable `death` is equals to the variable corresponding to the date of death in your dataset, then four analyses are performed:

- A first analysis censuring patients with no baseline, those with no follow-up and those who died without experiencing deterioration before;
- Sensitivity analysis ♯1: considering patients with no baseline and those with no follow-up score in deterioration;
- Sensitivity analysis ♯2: considering death as event;
- Sensitivity analysis ♯3: considering simultaneously patients with no baseline, those with no follow-up score and death as an event.

Arguments	Values
`X`	matrix or data frame
`score`	vector
`MCID`	scalar
`ref.init`	= `"baseline"` (default)/= `"best"` /= `"previous"`
`order`	= 1 (default) /= 2
`no_baseline`	= `"censure"` (default)/= `"event"`
`no_follow`	= `"censure"` (default)/= `"event"`
`death`	= `NA` (default) or vector
`sensitivity`	= `FALSE` (default)/= `TRUE`

Table 5: Arguments of the `TTD` function

The result of this function is a data frame with:

- the patient's identification number;
- a dummy variable called `event` equals to 1 if the patient is deteriorated, 0 if the patient is censored;
- a variable called `time` and equals to the time to censure or the time to the deterioration in months.

Like both variables `event` and `time` are created for each score treated and each definition of TTD, we added the name of the corresponding score as a suffix. As example, if `score = c("score1","score2")`. Then four variables are created: `event.score1`, `time.score1`, `event.score2` and `time.score2`.

Moreover, if `sensitivity == TRUE`, then added variables `event` and `time` are created:

- `event.SA1` for sensitivity analysis ♯1: `event` and `event.SA1` are equals except for patients with no baseline score and those with no follow-up score, `event.SA1=1` while `event = 0`. Like `time.SA1 == time`, then `time.SA1` was omitted;

- `event.SA2` and `time.SA2` for sensitivity analysis ♯2;
- `event.SA3` for sensitivity analysis ♯3. Like `time.SA3 == time.SA2`, then `time.SA3` was omitted.

- *Function* `TUDD()`

This function allows the estimation of the time until definitive deterioration according to the retained definition. All definitions of TUDD presented in this paper are implemented. The syntax of this function is nearly the same as for TTD. The next Table 6 summarizes the arguments of this function. Only one supplementary parameter as compared to `TTD()` function is available: the parameter `ref.def` in which you can specify the notion of "definitive deterioration" according to the three proposed definition in section 2 :

- With no further improvement of k points at least as compared to the reference score (`ref.def = "def1"`);
- with maintaining this deterioration of at least k points for all following scores, i.e., the deterioration is observed for all the following scores (`ref.def = "def2"`);
- with no further improvement of at least k points as compared to the score qualifying the deterioration (`ref.def = "def3"`).

Moreover, in this function, you can performed sensitivity analysis according to the MCID, thus the `MCID` parameter is a vector, not a scalar.

Arguments	Values
`X`	matrix or data frame with the data
`score`	vector
`MCID`	vector
`ref.init`	= `"baseline"` (default)/= `"best"` /= `"previous"`
`ref.def`	= `"def1"`/= `"def2"`/ = `"def3"`
`order`	= 1 (default) /= 2
`no_baseline`	= `"censure"` (default)/= `"event"`
`no_follow`	= `"censure"` (default)/= `"event"`
`death`	= `NA` (default) or vector
`sensitivity`	= `FALSE` (default)/= `TRUE`

Table 6: Arguments of the `TUDD` function

- *Function* `plotTTD()`

The package **QoLR** also contains a program called `plotTTD()` to obtain the TTD curves calculated according to the Kaplan-Meier estimation method for all patients or by treatment arm (only two groups are allowed). The `time` parameter is a vector equals to the time to deterioration or the time to censure and the `event` parameter is a dummy vector equals to 1 if the patient is deteriorated and 0 if not.

Some other information can be added in option like at regular time point `t` for all patients or by treatment arm:

- number of patients at risk (`nrisk=T`);
- cumulative number of events (`nevents=T`).

In the case of TTD curves by treatment arm you have to give the name of the group variable in the `group` parameter and the label of the each group you want to print in the `group.names` parameter, hazard ratio with 95% confidence interval and log-rank test can be added on the graph (`info = TRUE`) at a determined position specified by the user (`pos.info = c()`).

`xlab` and `ylab` correspond to the name of the abscissa and ordinate axis. The Table 7 summarizes the arguments of this function.

This function allows to plot the TTD curves with additional information useful for researchers to easily obtain standard curves for presentation or scientific publication.

Arguments	Values
`time`	vector
`event`	dummy vector
`group`	= `NULL` (default)/ vector
`nrisk`	= `TRUE` (default)/= `FALSE`
`nevent`	= `FALSE` (default)/= `TRUE`
`group.names`	= `NULL`/ vector
`t`	vector
`info`	= `FALSE` (default)/= `TRUE`
`pos.info`	= `NULL`/ vector
`xlab`	= character
`ylab`	= character

Table 7: Arguments of the `plotTTD` function

- *Function* `write.TTD()` and `write.TUDD()`

These tests can also be obtained for all the definitions of TTD or TUDD with `write.TTD()` and `write.TUDD()` respectively command available in **QoLR** package. These program create a Comma Separated Value file with the results of the TTD or TUDD analyses performed in one or more scores according to one main deterioration definition. All sensitivity analyses according to this primary definition are also performed, with one or more MCID, for all patients or by group (e.g., treatment arm effect). The results produced are:

- the number of patients initially at risk and the number of events;
- the median time of deterioration with 95% confidence interval;

and in the case analyses performed by group (only two groups are allowed):

- the results of log-rank test;
- the univariate Hazard Ratio with 95% confidence interval.

The arguments of this function is almost the same as for `TTD()` and `TUDD()` functions. Another parameter corresponds to the name of the file to print the results (`file = ""`).

4.2. Applications of the QoLR package

To apply one or more of these functions, you need to load the library **QoLR** in the R package with the next command:

```
R> library("QoLR")
```

We shall demonstrate the use of five main programs of **QoLR** in two datasets:

- one dataset `dataqol1` with the answers to the QLQ-C30 questionnaire for 20 patients;
- a second dataset `dataqol2` with data from longitudinal HRQOL measured on a 0-100 scale for 60 patients.

These datasets are available in **QoLR** package and can be imported in R via the command `data("dataqol1")` or `data("dataqol2")`.

Scoring EORTC questionnaires

We illustrated the use of the program `scoring.QLQC30()` to estimate the score of a QLQ-C30 questionnaire in the `dataqol1` dataset. Variables `q1` to `q30` are the answers to the 30 items of the QLQ-C30 for 20 patients. The first column of the data frame (`id`) is the patient's identification number. This dataset looks like the following:

```
R> data("dataqol1")
R> head(dataqol1)
```

	id	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13	q14	q15	q16	q17	q18	q19	q20
1	1	3	3	2	3	1	1	4	1	1	1	2	1	2	3	1	3	1	1	1	1
2	2	1	1	1	1	1	1	1	1	2	1	3	2	1	2	1	2	1	1	1	1
3	3	2	NA	1	1	1	1	2	1	2	NA	4	2	1	2	4	1	2	NA	3	NA
4	4	1	3	1	1	3	2	2	NA	1	2	2	2	1	1	1	2	2	2	1	1
5	5	1	3	1	1	NA	2	2	3	1	2	2	NA	1	1	1	1	1	2	1	1
6	6	3	2	1	1	1	1	1	2	1	2	3	2	1	1	1	1	1	2	1	1

	q21	q22	q23	q24	q25	q26	q27	q28	q29	q30
1	2	1	1	1	2	3	3	1	5	5
2	2	2	2	3	2	2	1	1	6	6
3	1	2	1	1	1	1	1	1	NA	NA
4	1	1	1	1	1	1	1	1	5	5
5	1	1	1	1	1	1	2	1	5	5
6	1	1	1	1	1	1	1	1	5	6

To apply the `scoring.QLQC30()` program, you have to specify at least the name of the dataset with the items answers and the items position. By default, items position is `1:30`. You can also specify a patient's identification number in the `id` parameter. The result is a data frame `score_base` in which each variable corresponds to a score of the questionnaire and the first variable is the `id` parameter if it is filled:

```
R> score_base=scoring.QLQC30(dataqol1, id="id", items=2:31)
R> head(round(score_base))
```

```
  id QL2 PF2 RF2  EF  CF  SF FA NV PA DY  SL AP CO DI FI
1  1  67  53  50  92  83  33  0 33  0  0  33 33 67  0  0
2  2  83 100 100  58  83  83 11 17 17  0  67  0 33  0  0
3  3  NA  92  83  92 100 100 NA 67 50  0 100  0  0 33  0
4  4  67  73  67 100 100 100 33  0  0 NA  33  0 33 33  0
5  5  67  83  67 100 100  83 33  0  0 67  33  0  0  0  0
6  6  75  80 100 100 100 100 33  0  0 33  67  0  0  0  0
```

Time to HRQOL score deterioration

- Time to deterioration

In order to illustrate the use of the programs `TTD()` and `TUDD()` for the estimation of the TTD in a HRQOL score, we used a second dataset `dataqol2`. This dataset contains HRQOL measures obtained on a visual analogue scale from 0 to 100 with 6 measures per patient. Two assessments were done at each measurement time:

- a global HRQOL assessment (`QoL`);
- a pain intensity assessment (`pain`).

A high score for `QoL` corresponds to a high HRQOL level and a high level for `pain` corresponds to a high presence of pain. Some scores and/or dates are missing to illustrate patients profile with no baseline or no follow up as well as intermittent and monotone missing values treatment. This table also contains the date of death for patients died during the study (`death` variable). Patients in this table were randomly allocated to one treatment group corresponding to the variable `arm` (dichotomous variable equals to 0 or 1). This dataset is available in **QoLR** package via the command `data("dataqol2")` and it looks like the following:

```
R> head(dataqol2)
```

```
  id time date QoL pain arm death
1  1    0    0  78   36   0   201
2  1    1   49  40   46   0   201
3  1    2  102  60   32   0   201
4  1    3  145  54   25   0   201
5  1    4  201  63   26   0   201
6  1    5  232  48   33   0   201
```

In this dataset, the `id` variable is the patient's identification number. The dataset is sorted by patient's identification number (`id`) and HRQoL measures (`time`). The variable `time` corresponds to the theoretical measure of HRQOL. As example, `time = 0` corresponds to the baseline measure in this dataset; `time = 1` corresponds to the second assessment of HRQOL.

The variable `date` corresponds to the date of HRQOL measure in Julian format. The baseline date was set to 0 and all the next dates correspond to the time between the measure and the baseline date (the baseline date is the date of origin). HRQOL was evaluated around every 50 days. In the same way, `death` corresponds to the time between baseline date and the date of death if the patient died during the study. The HRQOL measures correspond to the variable `QoL` for HRQOL global score and `pain` for pain score.

Like HRQOL was measured at several measurement times for each patient, we can study the time to deterioration of the HRQOL scores. To begin, we can study the reference definition of TTD of the score `QoL`, i.e., a deterioration as compared to the baseline score (`ref.init = "baseline"`) with at least 5-point MCID (`MCID = 5`) and considering patients with no baseline or with no follow-up measure censoring at baseline or just after baseline (`no_baseline = "censure"` and `no_follow = "censure"`). The score `QoL` corresponds to a measure of global HRQOL. In this way, a deterioration corresponds to a decrease of the score (`order = 1`). Like the values of these parameters are the default values except for the MCID, we do not need to specify their values and the function can be applied as follows:

```
R> ttd1=TTD(dataqol2, score="QoL", MCID=5)
R> head(ttd1)

  id event.QoL   time.QoL
1  1         1 1.60985626
2  2         0 0.03285421
3  3         0 0.00000000
4  4         1 1.80698152
5  5         1 5.09240246
6  6         1 6.80082136
```

The result is a data frame with the identification number of patients (`id`), the time to deterioration or to censure in months (`time.QoL`) and a dummy variable (`event.QoL`) indicating if the patient is deteriorated (`event.QoL =1`) or not (`event.QoL =0`). The suffix "QoL" corresponds to the name of the treated score. According to this definition, 44 patients are deteriorated among the 60 patients regarding the `QoL` score:

```
R> sum(ttd1$event.QoL,na.rm=T)

[1] 44
```

If we want to consider patients with no baseline or no follow up as events, we have to fix the parameters `no_baseline` and `no_follow` to `"event"` as follows:

```
R> ttd2=TTD(dataqol2, score="QoL", order=1, MCID=5, no_baseline="event",
+ no_follow="event")
R> head(ttd2)

 id event.QoL   time.QoL
1  1         1 1.60985626
```

```
2  2         1 0.03285421
3  3         1 0.00000000
4  4         1 1.80698152
5  5         1 5.09240246
6  6         1 6.80082136
```

```
R> sum(ttd2$event.QoL,na.rm=T)
```

```
[1] 50
```

6 patients with no baseline or no follow-up measure are then added to the events.

To consider death as an event, we have to specify the value of the `death` parameter:

```
R> ttd3=TTD(dataqol2, score="QoL", MCID=5, death="death")
R> head(ttd3)
```

```
  id event.QoL   time.QoL
1  1         1 1.60985626
2  2         0 0.03285421
3  3         0 0.00000000
4  4         1 1.80698152
5  5         1 5.09240246
6  6         1 6.80082136
```

Finally, we can obtain directly all the sensitivity analyses along with the primary analysis in one application of the function `TTD` by specify `sensitivity = TRUE`:

```
R> ttd4=TTD(dataqol2, score="QoL", MCID=5, death="death", sensitivity=TRUE)
R> head(ttd4)
```

```
  id event.QoL   time.QoL event.SA1.QoL event.SA2.QoL time.SA2.QoL event.SA3.QoL
1  1         1 1.60985626             1             1   1.60985626             1
2  2         0 0.03285421             1             0   0.03285421             1
3  3         0 0.00000000             1             0   0.00000000             1
4  4         1 1.80698152             1             1   1.80698152             1
5  5         1 5.09240246             1             1   5.09240246             1
6  6         1 6.80082136             1             1   6.80082136             1
```

When all sensitivity analyses are performed, some added variables are created. Variables `event.QoL` and `time.QoL` still correspond to the results for the primary analysis (TTD as compared to the baseline score in the present case). `event.SA1.QoL` is the dummy event variable considering patients with no baseline or no follow up measure in deterioration since baseline, while they were censured in the primary analysis (Sensitivity Analysis ♯1). Like the corresponding times are the same as in the primary analysis, no new time variable was created. Variables `event.SA2.QoL` and `time.SA2.QoL` are the results for the analysis adding death as

an event (Sensitivity Analysis ♯2). Finally, `event.SA3.QoL` corresponds to the event variable with death and patients with no baseline or no follow up as events (Sensitivity Analysis ♯3).

In order to integrate the response shift effect, we can choose the best previous HRQOL score or the previous score (i.e., immediately preceding score) as reference score by specifying `ref.init="best"` or `ref.init="previous"` respectively.

In the following example, the reference score is the best previous `QoL` score:

```
R> ttd5=TTD(dataqol2, score="QoL", MCID=5, ref.init ="best")
R> head(ttd5)

  id event.QoL   time.QoL
1  1         1 1.60985626
2  2         0 0.03285421
3  3         0 0.00000000
4  4         1 1.80698152
5  5         1 5.09240246
6  6         1 4.46817248

R> sum(ttd5$event.QoL,na.rm=T)

[1] 54
```

Then, 54 patients experienced a deterioration as compared to the best previous `QoL` score

The function `TTD()` can handle simultaneously many scores, functional and/or symptomatic. You have to define scores names studied in the `score` parameter and the order to considered (decrease or increase): `order = 1` for the functional scale `QoL` and `order = 2` for the symptomatic scale `pain`. Variables `event` and `time` are then created for each score with the score name as a suffix. The following example represents the application of the TTD as compared to the baseline score with a 5-point MCID for `QoL` and `pain` respectively:

```
R> ttd6=TTD(dataqol2, score=c("QoL", "pain"), order=1:2, MCID=5)
R> head(ttd6)

  id event.QoL   time.QoL event.pain  time.pain
1  1         1 1.60985626          1 1.60985626
2  2         0 0.03285421          0 0.03285421
3  3         0 0.00000000          0 0.00000000
4  4         1 1.80698152          1 1.80698152
5  5         1 5.09240246          1 1.60985626
6  6         1 6.80082136          1 1.90554415
```

- Time until definitive deterioration

The TUDD is studied with the function `TUDD()`, quite similar to the function `TTD()`. By default, the deterioration corresponds to a deterioration with a k-point MCID as compared

with the baseline score with no further improvement of k points at least as compared to the baseline score (Bonnetain *et al.* 2010). The result of the application of this function is fairly similar to the one of the `TTD()`. However, for `TUDD`, the value of the MCID is also specified in the variables name `time` and `event`. The result of the reference definition of TUDD is the following:

```
R> tudd1=TUDD(dataqol2, score="QoL", MCID=5)
R> head(tudd1)
```

```
  id event.5.QoL time.5.QoL
1  1           1 1.60985626
2  2           0 0.03285421
3  3           0 0.00000000
4  4           0 7.81930185
5  5           1 5.09240246
6  6           1 6.80082136
```

```
R> sum(tudd1$event.5.QoL,na.rm=T)
```

```
[1] 34
```

Thus, 34 patients experienced a definitive deterioration as compared to the baseline score according to the definition of Bonnetain *et al* (Bonnetain *et al.* 2010). The deterioration can also be definitive as compared to the deterioration observed, i.e., with no further improvement of 5-point MCID as compared to the score obtained at the time of the first deterioration. This definition is applied by setting the parameter `ref.def` to the value `"def3"`.

```
R> tudd2=TUDD(dataqol2, score = "QoL", MCID = 5, ref.def = "def3")
R> head(tudd2)
```

```
  id event.5.QoL time.5.QoL
1  1           1 3.35112936
2  2           0 0.03285421
3  3           0 0.00000000
4  4           0 7.81930185
5  5           1 5.09240246
6  6           1 6.80082136
```

Like we did for the TTD, all the sensitivity analyses can be performed simultaneously with the primary definition of TUDD. Moreover, many MCID can be specified. In fact, as it was defined in section 2, we need a dependence between sensitivity analyses varying the MCID for TUDD. An indicator of the MCID value is added as a suffix of the resulting parameters `event` and `time`.

```
R> tudd3=TUDD(dataqol2, score="QoL", MCID=c(5,10), sensitivity=T)
R> head(round(tudd3,2))
```

```
 id event.10.QoL time.10.QoL event.10.SA1.QoL event.5.QoL time.5.QoL event.5.SA1.QoL
1  1            1       1.61                1           1      1.61                1
2  2            0       0.03                1           0      0.03                1
3  3            0       0.00                1           0      0.00                1
4  4            0       7.82                0           0      7.82                0
5  5            0       6.74                0           1      5.09                1
6  6            1       6.80                1           1      6.80                1
```

In this application, death has not been taken into account. Only two sensitivity analyses were performed: one on the MCID value and the second one regarding patients with no baseline or no follow-up measure.

- Time to deterioration curves

The Figure 2 corresponds to the TUDD of `QoL` score as compared to the baseline score with a 5-point MCID according treatment arm (arm effect). In this graph, we printed the number of patients still at risk at each time point according to treatment arm (`nrisk=T`). Moreover, the result of the log-rank test and the hazard ratio of arm 2 vs. arm 1 is also printed (`info=T`, `pos.info=c(5,0.8)`). The hazard ratio (arm 2 vs. arm 1) equals to 1.20 with 95% confidence interval $(0.69 - 2.09)$ and the result of the log-rank test is $p = 0.523$.

```
R> tudd1=TUDD(dataqol2, score="QoL", MCID=5,ref.init="baseline",ref.def="def1")
R> ttd_1=merge(tudd1,unique(dataqol2[,c("id","arm")]))
R> plotTTD(ttd_1$time.5.QoL,ttd_1$event.5.QoL,ttd_1$arm,nrisk=T,nevent=F,
+ group.names=c("arm 1","arm 2"), t=seq(0,8,2),info=T,pos.info=c(6,0.8),
+ xlab="time (months)", ylab="probability (%)")
```

- Write all the results of a time to deterioration analysis in a csv file

These tests (hazard ratio and log-rank test) and other information can also be obtained for all the definitions of TTD or TUDD with `write.TTD()` or `write.TUDD()` commands available in **QoLR** package.

As example, the following command creates a csv file named `"file_TTD_baseline"` which is located in the current directory of your R session. Table 8 is an extraction of this file for 5 points MCID.

```
R> write.TTD(dataqol2, score=c("QoL","pain"), order=c(1,2), MCID=c(5,10),
+ group="arm", names.group=c("arm 1","arm 2"), sensitivity=FALSE,
+ file="file_TTD_baseline")
```

As example, in arm 1 and arm 2 respectively, 21 and 23 patients experienced a definitive deterioration of `QoL` score of 5 points at least as compared to the baseline score, among the 60 patients included (30 patients per arm). The median TUDD was 5.01 months with a 95% confidence interval $(3.25 - NA)$ for arm 1 and 3.30 months with a 95% confidence interval $(1.91 - 5.13)$ for arm 2 (log-rank $P = 0.134$). The univariate Cox Hazard Ratio of arm 2 vs. arm 1 was 1.57 with a 95% confidence interval $(0.87 - 2.85)$.

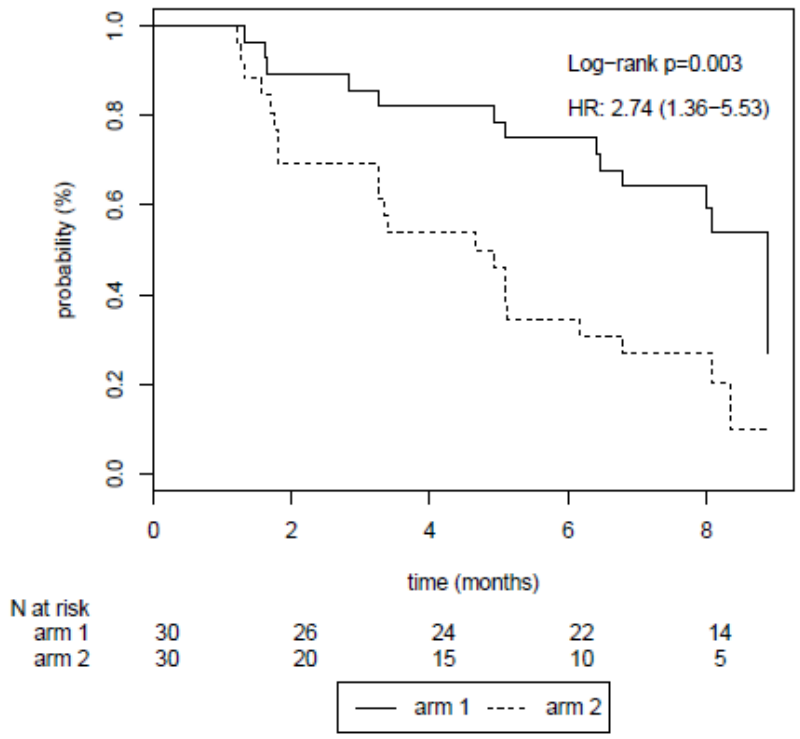


Figure 2: Kaplan-Meier survival curve for the Time until definitive QoL score deterioration of at least 5-point MCID as compared to the baseline score

	MCID = 5 points TTD baseline n (events)	median (CI 95%)	Log-rank	HR (CI 95%)
QoL				
arm 1	30 (21)	5.01 (3.25-NA)	p=0.134	1
arm 2	30 (23)	3.3 (1.91-5.13)		1.57 (0.87-2.85)
pain				
arm 1	30 (23)	1.76 (1.61-2)	p=0.493	1
arm 2	30 (19)	1.77 (1.64-NA)		0.81 (0.44-1.49)

Table 8: The csv file created with the application of the `write.TTD` function

5. Conclusion and outlook

The **QoLR** package is the first R package dedicated to the analysis of HRQOL. The implementation of the time to deterioration definitions in a HRQOL score allows the dissemination of

these approaches in order to reach the goal of standardization of longitudinal HRQOL studies.

QoLR will be updated as new modules will be developed by the EORTC HRQOL group. The package will be completed over time, by some simulations algorithms of longitudinal HRQOL data with intermittent or monotone missing data of type Missing Completely At Random or Missing Not At Random for example. Other programs will allow to print the results of Univariate and Multivariate Cox analyses in a csv file. Moreover, an ongoing project investigates the presence of competitive risk between events. This package will then be complemented with some competitive risk models. Finally, the time to deterioration approach is currently implemented under the SAS software (SAS Institute Inc 2011) to allow researchers not familiar with R software to apply this longitudinal analysis method.

Acknowledgements

This study was supported by a grant from the French Public Health Research Institute (http://IRESP.net) under the 2012 call for projects as part of the 2009-2013 Cancer Plan.

References

Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JC, *et al.* (1993). "The European Organization for Research and Treatment of Cancer QLQ-C30: A Quality-of-life Instrument for Use in International Clinical Trials in Oncology." *Journal of the National Cancer Institute*, **85**(5), 365–376.

Ahmed S, Mayo NE, Corbiere M, Wood-Dauphinee S, Hanley J, Cohen R (2005). "Change in Quality of Life of People with Stroke Over Time: True Change or Response Shift?" *Quality of Life Research*, **14**(3), 611–627.

Anota A, Hamidou Z, Paget-Bailly S, Chibaudel B, Bascoul-Mollevi C, Auquier P, Westeel V, Fiteni F, Borg C, Bonnetain F (2013). "Time to Health-Related Quality of Life Score Deterioration as a Modality of Longitudinal Analysis for Health-Related Quality of Life Studies in Oncology: Do We Need RECIST for Quality of Life to Achieve Standardization?" *Quality of Life Research*, pp. 1–14.

Bonnetain F, Dahan L, Maillard E, Ychou M, Mitry E, Hammel P, Legoux JL, Rougier P, Bedenne L, Seitz JF (2010). "Time Until Definitive Quality of Life Score Deterioration as a Means of Longitudinal Analysis for Treatment Trials in Patients with Metastatic Pancreatic Adenocarcinoma." *European Journal of Cancer*, **46**(15), 2753–2762.

Bullinger M (2002). "Assessing Health Related Quality of Life in Medicine. An Overview over Concepts, Methods and Applications in International Research." *Restorative Neurology and Neuroscience*, **20**(3), 93–101.

Cnaan A, Laird N, Slasor P (1997). "Tutorial in Biostatistics: Using the General Linear Mixed Model to Analyse Unbalanced Repeated Measures and Longitudinal Data." *Statistics in Medicine*, **16**, 2349–2380.

Diggle PJ (1988). "An Approach to the Analysis of Repeated Measurements." *Biometrics*, pp. 959–971.

Fairclough DL (2010). *Design and Analysis of Quality of Life Studies in Clinical Trials*. CRC press.

Fayers PM, Aaronson NK, Bjordal K, Curran D, Grønvold M (1999). "EORTC QLQ-C30 Scoring Manual." *Technical report*, EORTC.

Gibbons F (1999). "Social Comparison as a Mediator of Response Shift." *Social Science & Medicine*, **48**(11), 1517–1530.

Goel MK, Khanna P, Kishore J (2010). "Understanding Survival Analysis: Kaplan-Meier Estimate." *International Journal of Ayurveda Research*, **1**(4), 274.

Hamidou Z, Dabakuyo TS, Mercier M, Fraisse J, Causeret S, Tixier H, Padeano MM, Loustalot C, Cuisenier J, Sauzedde JM, *et al.* (2011). "Time to Deterioration in Quality of Life Score as a Modality of Longitudinal Analysis in Patients with Breast Cancer." *The Oncologist*, **16**(10), 1458–1468.

Hunger M, Döring A, Holle R (2012). "Longitudinal Beta Regression Models for Analyzing Health-Related Quality of Life Scores Over Time." *BMC Medical Research Methodology*, **12**(1), 144.

Korfage IJ, de Koning HJ, Essink-Bot ML (2007). "Response Shift due to Diagnosis and Primary Treatment of Localized Prostate Cancer: A Then-test and a Vignette Study." *Quality of Life Research*, **16**(10), 1627–1634.

Osoba D (2011). "Health-Related Quality of Life and Cancer Clinical Trials." *Therapeutic Advances in Medical Oncology*, **3**(2), 57–71.

Osoba D, Rodrigues G, Myles J, Zee B, Pater J (1998). "Interpreting the Significance of Changes in Health-Related Quality-of-Life Scores." *Journal of Clinical Oncology*, **16**(1), 139–144.

Pan AW, Chen YL, Chung LI, Wang JD, Chen TJ, Hsiung PC (2012). "A Longitudinal Study of the Predictors of Quality of Life in Patients with Major Depressive Disorder Utilizing a Linear Mixed Effect Model." *Psychiatry Research*, **198**(3), 412–419.

Pauler DK, McCoy S, Moinpour C (2003). "Pattern Mixture Models for Longitudinal Quality of Life Studies in Advanced Stage Disease." *Statistics in Medicine*, **22**(5), 795–809.

Penar-Zadarko B, Binkowska-Bury M, Wolan M, Gawelko J, Urbanski K (2012). "Longitudinal Assessment of Quality of Life in Ovarian Cancer Patients." *European Journal of Oncology Nursing*.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

SAS Institute Inc (2011). *The SAS System, Version 9.3*. Cary, NC. URL http://www.sas.com/.

Sprangers MA, Schwartz CE (1999). "Integrating Response Shift into Health-Related Quality of Life Research: A Theoretical Model." *Social Science & Medicine*, **48**(11), 1507–1515.

Therneau TM (2014). *A Package for Survival Analysis in S*. R package version 2.37-7, URL http://CRAN.R-project.org/package=survival.

Ubel PA, Peeters Y, Smith D (2010). "Abandoning the Language of "Response Shift": A Plea for Conceptual Clarity in Distinguishing Scale Recalibration from True Changes in Quality of Life." *Quality of Life Research*, **19**(4), 465–471.

Wiklund I (2004). "Assessment of Patient-Reported Outcomes in Clinical Trials: The Example of Health-Related Quality of Life." *Fundamental & Clinical Pharmacology*, **18**(3), 351–363.

Zeileis A, Grothendieck G (2005). "**zoo**: S3 Infrastructure for Regular and Irregular Time Series." *Journal of Statistical Software*, **14**(6), 1–27. URL http://CRAN.R-project.org/package=zoo.

Affiliation:

Amélie Anota
Quality of Life in Oncology clinical research Platform
Methodology and Quality of Life in Oncology Unit (EA 3181)
University Hospital of Besançon
25 030 Besançon Cedex
France
E-mail: aanota@chu-besançon.fr

Marion Savina
INSERM
Clinical and Epidemiological Research Unit (CIC-EC 7) – CTD INCa
Institut Bergonie
33 000 Bordeaux
France
E-mail: Marion.Savina@isped.u-bordeaux2.fr

Caroline Bascoul-Mollevi
Biostatistics Unit
Cancer Institute
34 090 Montpellier Val d'Aurelle
France
E-mail: Caroline.Mollevi@icm.unicancer.fr

Franck Bonnetain
Quality of Life in Oncology clinical research Platform
Methodology and Quality of Life in Oncology Unit (EA 3181)
University Hospital of Besançon
25 030 Besançon Cedex
France
E-mail: franck.bonnetain@univ-fcomte.fr

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd

2.3. TJD et impact des données manquantes aléatoires : investigation du score de propension pour tenir compte des données manquantes

Résumé

Objectif

Le TJD est une méthode d'analyse longitudinale de la QdV régulièrement utilisée dans les essais cliniques en cancérologie (Bonnetain *et al*, 2010; Gourgou-Bourgade *et al*, 2013). Dans la définition du TJD, les données manquantes au cours du suivi sont ignorées, considérant que le niveau de QdV du patient est resté constant depuis la dernière mesure disponible. Or, les données manquantes sont souvent liées soit au niveau de QdV non observée du patient (type MNAR), soit à des caractéristiques du patient observées à l'inclusion par exemple (type MAR). Il est nécessaire de tenir compte de ces données manquantes dans la méthode d'analyse longitudinale.

Le score de propension est souvent utilisé dans les études non randomisées afin de réduire le biais dû à cette absence de randomisation (Trojano *et al*, 2009). Différentes méthodes basées sur le score de propensions existent ; la méthode « inverse probability of treatment weighting » semble la plus adéquate pour les modèles de survie (Austin, 2013).

L'objectif était d'investiguer la méthode du temps jusqu'à détérioration (TJD) d'un score de qualité de vie relative à la santé (QdV) comme modalité d'analyse longitudinale dans un essai de phase II sur le cancer du pancréas en utilisant la méthode « inverse probability of treatment weighting » du score de propension pour tenir compte des données manquantes dépendantes des caractéristiques des patients (données manquantes de type MAR).

Matériels et méthodes

Un essai de phase II multicentrique randomisé a été réalisé pour évaluer la stratégie d'un traitement séquentiel par FOLFIRI.3 + gemcitabine versus gemcitabine seule, chez des patients atteints d'un cancer du pancréas métastatique non prétraité.

La QdV a été évaluée par le questionnaire EORTC QLQ-C30 à l'inclusion, puis tous les deux mois jusqu'à arrêt de l'étude ou le décès.

Les profils de patients constitués étaient les patients avec au moins un score manquant lors du suivi (répondeurs incomplets) versus ceux avec tous les scores disponibles jusqu'à leur sortie d'étude ou le décès (répondeurs complets). Ces profils ont été comparés par un modèle de

régression logistique multivarié intégrant toutes les variables recueillies à l'inclusion et significatives en univarié au seuil $P \leq 0.20$. Les valeurs prédites par le modèle ont ensuite été extraites pour constituer le score de propension.

L'analyse longitudinale des données de QdV a été réalisée selon la méthode du TJD. Le TJD a été défini comme le temps entre la randomisation et l'observation d'une première détérioration d'au moins 5 points par rapport au score à l'inclusion sans amélioration ultérieure de plus de 5 points, ou le décès (Bonnetain *et al*, 2010). Des analyses de Cox ont été réalisées afin d'identifier les facteurs influençant le TJD. Pour tenir compte du profil MAR des données manquantes, les analyses ont été reprises en pondérant selon la méthode « inverse probability of treatment weighting » du score de propension (Austin, 2013). Ainsi, une pondération a été appliquée aux patients selon la présence ou non des données manquantes.

Résultats

98 patients ont été inclus entre octobre 2007 et mai 2011. Parmi les 66 patients ayant répondus aux questionnaires de QdV (67.3%), 40 étaient des répondeurs complets (60.6%) et 26 des répondeurs complets (39.4%) durant le suivi.

Le modèle de régression logistique multivarié a mis en évidence une association entre le profil de répondeurs partiels et un site de la tumeur primitive : la tête (Odds Ratio (OR) = 2.72 [0.86-9.16]), la présence de métastases au niveau des lymphocytes (OR = 7.90 [1.12-164.12]) et un taux d'hémoglobine élevé (OR = 1.80 [0.54-6.10]).

Concernant les résultats de l'analyse longitudinale du TJD sans pondération, les patients du bras FOLFIRI.3 + gemcitabine présentaient un TJD significativement plus long que ceux du bras gemcitabine seule pour le fonctionnement physique (Hazard Ratio (HR) = 0.40 (CI 95% 0.20-0.82)) et la douleur (HR = 0.47 (0.23-0.98)). Après pondération, les patients du bras FOLFIRI.3+gemcitabine présentaient également un TJD significativement plus long que ceux du bras gemcitabine seule pour la santé globale (HR : 0.52 [IC 95% 0.31-0.85]), le fonctionnement émotionnel (HR : 0.35 [0.21-0.59]) et la fatigue (HR : 0.61 [0.38-0.97]).

Conclusion

Les patients du bras FOLFIRI.3+gemcitabine présentent globalement un TJD plus long que ceux du bras gemcitabine. Le TJD a l'avantage de fournir des résultats significatifs pour les cliniciens. De plus, la méthode « inverse probability of treatment weighting » du score de propension permet de prendre en compte le profil MAR des données manquantes.

<u>Article:</u> Sequential FOLFIRI.3 + Gemcitabine improves health-related quality of life deterioration-free survival of patients with metastatic pancreatic adenocarcinoma: a randomized phase II trial

Article en révision dans *PLOS ONE*

Short title: **Quality of life in metastatic pancreatic adenocarcinoma**

A. Anota[1,2], G. Mouillet[2], I. Trouilloud[3], A.C. Dupont-Gossart[4], P. Artru[5], T. Lecomte[6], A. Zaanan[3], M. Gauthier[7], F. Fein[4], O. Dubreuil[3], S. Paget-Bailly[2], J. Taieb[3], F. Bonnetain[1,2]

[1] Quality of Life in Oncology National Platform, France

[2] Methodological and Quality of Life in Oncology Unit, EA 3181, University Hospital of Besançon, Besançon, France

[3] Department of Gastroenterology and Digestive Oncology, Georges Pompidou European Hospital, University of Paris Descartes, Paris, France

[4] Department of Gastroenterology, University Hospital of Besançon, Besançon, France

[5] Hepato-Gastro-Enterology and Digestive Oncology Department, Hospital Jean Mermoz, Lyon, France

[6] Department of Gastroenterology and Digestive Oncology, CHU de Tours-Hopital Trousseau, Chambray-Les-Tours, France

[7] Biostatistics and Quality of life unit, Centre Georges François Leclerc, Dijon, France

Author Contributions:

Study concept: IT, AC. DG, PA, TL, AZ, FF, OD, JT, FB

Study design: IT, AC. DG, PA, TL, AZ, FF, OD, JT, FB

Data acquisition: IT, AC. DG, PA, TL, AZ, FF, OD, JT, FB

Quality control of data and algorithm: MG, FB, AA

Statistical analysis: AA, FB, MG

Manuscript preparation: AA, GM, S P-B, FB

Manuscript editing: all authors

Manuscript review: all authors

Corresponding author:

Amélie Anota

Quality of Life in oncology clinical research platform

Methodological and Quality of Life in Oncology Unit (EA 3181)

University Hospital of Besançon

France

Telephone number: +33381218896

Fax number: +33381665299

Email: aanota@chu-besancon.fr

Abstract

Background

A randomized multicenter phase II trial was conducted to assess the sequential treatment strategy using FOLFIRI.3 and gemcitabine alternately (Arm 2) compared to gemcitabine alone (Arm 1) in patients with metastatic non pre-treated pancreatic adenocarcinoma. The primary endpoint was the progression-free survival (PFS) rate at 6 months. It concludes that the sequential treatment strategy appears to be feasible and effective with a PFS rate of 43.5% in Arm 2 at 6 months (26.1% in Arm 1). This paper reports the results of the longitudinal analysis of the health-related quality of life (HRQoL) as a secondary endpoint of this study.

Methods

HRQoL was evaluated using the EORTC QLQ-C30 at baseline and every two months until the end of the study or death. HRQoL deterioration-free survival (QFS) was defined as the time from randomization to a first significant deterioration as compared to the baseline score with no further significant improvement, or death. A propensity score (PS) was estimated comparing characteristics of partial and complete responders. Analyses were repeated with inverse probability weighting method using the PS. Multivariate Cox regression analyses were performed to identify independent factors influencing QFS.

Results

98 patients were included between 2007 and 2011. Adjusting on the PS, patients of Arm 2 presented a longer QFS of Global Health Status (Hazard Ratio: 0.52 [0.31-0.85]), emotional functioning (0.35 [0.21-0.59]) and pain (0.50 [0.31 – 0.81]) than those of Arm 1.

Conclusion

Patients of Arm 2 presented a better HRQoL with a longer QFS than those of Arm 1. Moreover, the PS method allows to take into account the missing data depending on patients' characteristics.

Keywords: Health-related quality of life, oncology clinical trials, longitudinal analysis, time until definitive deterioration, missing data, propensity score

Abbreviations

CI: confidence interval

EORTC: European Organisation for Research and Treatment of Cancer

HR: Hazard Ratio

HRQoL: health-related quality of life

IPW: inverse probability weighting

MCID: minimal clinically important difference

MD: missing data

mPC: metastatic Pancreatic Cancer

OR: Odds-Ratio

PS: propensity score

QFS: HRQoL deterioration-free survival

SD: standard deviation

TUDD: time until definitive deterioration

INTRODUCTION

The results of a phase II trial concerning untreated patient with metastatic Pancreatic Cancer (mPC) have shown that sequential treatment using FOLFIRI.3 and gemcitabine was effective and safe [1].

In first line treatment, FOLFIRINOX protocol and the association of nab-paclitaxel + gemcitabine improve overall survival (OS) [2,3] and represent a new therapeutic option in first line. However, the less favorable toxicity profiles of these new strategies could limit this option to younger patients with a good PS (0 or 1) [4]. A sequential association of chemotherapy protocol without cross-resistance may increase anti-tumor effects and limit toxicities, preserving patient's Health-related Quality of Life (HRQoL).

Prognosis of patients with mPC remains extremely poor. In consequence, HRQoL is a major subject of concern for these patients who are often painful and symptomatic at the time of diagnosis. Moreover, HRQoL appears to be an independent prognostic factor for OS alongside classical clinical and demographic factors [5]. In metastatic settings, the current discussion is to consider HRQoL as a co-primary endpoint along with a tumor parameter such as progression-free survival (PFS) [6,7].
However, HRQoL results remain poorly used to modify therapeutic strategies, due to the complexity of its longitudinal analysis and to a lack of standardization. Moreover, results should have the ability to translate findings into information that decision makers find understandable and compelling.
In recent years, time to event models like time until definitive HRQoL score deterioration (TUDD) have been proposed as a modality of longitudinal HRQoL analysis in oncology, especially in metastatic setting [8]. The TUDD method produces clinically meaningful results for clinicians like Kaplan-Meier survival curves and hazard ratio (HR). TUDD including death as an event was defined as "HRQoL deterioration-free survival" (QFS) [9].

One other major concern of longitudinal HRQoL studies is missing data (MD) [10], specifically in advanced cancer where attrition is common [11]. Patients may dropout before the end of the study, generally due to a health status deterioration or death. In this case, MD can bias the analysis and interpretation [10,12,13,14], and should be considered to ensure accuracy and robustness of the results. Several methods have

been investigated to handle with MD [15,16]. The most well-known is the pattern-mixture model [17] but it is rarely applied due to its complexity [17,18].
Then it would be interesting to develop a method to use in conjunction with QFS to handle with informative MD. Methods using the propensity score (PS) are often used in observational studies in order to reduce the bias of the absence of randomization and to allow causal inference [19]. The PS is used to model the probability of receiving a treatment conditionally to the variables observed before treatment. The main methods used with PS are stratification, matching and inverse probability weighting (IPW) methods [20]. In survival analyses, IPW method is recommended [21].

The objective of this study was to compare longitudinal HRQoL according to treatment arm using QFS in a metastatic setting and secondary to investigate the application of the IPW method based on the PS in conjunction with the TUDD in order to take into account MD depending on patients' characteristics.

MATERIALS AND METHODS

Patients and eligibility criteria

This study was a multicenter, randomized, non-comparative, open phase II trial, conducted in French centers. Inclusion criteria were: histologically or cytologically proven mPC, no previous chemotherapy (adjuvant chemotherapy with gemcitabine was allowed if administered more than 12 months before inclusion) or radiotherapy (unless at least one measurable target lesion was present outside the irradiated area) and WHO performance status <2. Exclusion criteria were bile ducts adenocarcinoma, ampulloma and a history of another cancer. All patients were fully informed of the study and provided signed written informed consent. The protocol was approved by the ethics committees ("Comité de Protection des Personnes"). This study FIRGEM was registered with EudraCT (https://eudract.ema.europa.eu/; N° 2006-005703-34) before the start date. The design of this study has been extensively described elsewhere [1] .

Using minimization technique, patients were randomly (ratio 1:1) assigned to receive sequentially FOLFIRI.3 every 14 days during two months (four courses per cycle), followed by gemcitabine (6 courses at days 1, 8, 15, 29, 36 and 43 per cycle) (Arm 2) or gemcitabine alone (Arm 1). A deterministic minimization was employed and stratification criteria were center (10 centers), performance status (0 vs. 1) and the number of metastatic sites (one vs. more than one).

Health-related Quality of Life assessment

HRQoL was evaluated using the European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30 cancer specific questionnaire [22], at inclusion and every two months until progression, limiting toxicity, patient's refusal or death. The QLQ-C30 includes 30 items and measures five functional scales (physical, role, emotional, cognitive and social functioning), global health status (GHS), financial difficulties and eight symptom scales (fatigue, nausea and vomiting, pain, dyspnea, insomnia, appetite loss, constipation, diarrhea) [22]. These scores vary from 0 (worst) to 100 (best) for the functional dimensions and GHS, and from 0 (best) to 100 (worst) for the symptom dimensions and were generated according to the EORTC Scoring Manual [23].

Statistical analysis

Sample size calculation

The primary endpoint was the 6-month PFS rate. Secondary endpoints were OS, safety/tolerability, tumor response, PFS and QFS. The trial was based on a Fleming one-step design [24]. The expected 6-month PFS rate with the sequential treatment was 45%. A PFS rate of 25% was chosen as uninteresting rate of effectiveness (H0: 6-month PFS 25% = unacceptable efficacy, H1: 6-month PFS 45% = expected efficacy). With a unilateral type I error of 5% and a type II error of 10%, it was necessary to include 46 patients in each arm, rounded to 49 to compensate for an anticipated 5% rate of loss to follow-up.

Based on Fleming decision criteria, experimental arm will be considered uninteresting if 15 or less than 15 alive patients were free of progression. It will be considered as promising if 16 or more than 16 alive patients were free of progression.

The analysis was performed on intent-to-treat principle (all randomized patients irrespective of treatment received and eligibility criteria). Analyses of primary endpoint were done on the first randomized 46 patients with available PFS data (to match with Fleming criteria decision rules) while all other analyses were done on all randomized patients.

Population

Randomized patients whatever eligibility criteria with at least one HRQoL score were included in the QFS analysis (modified intent to treat analysis). Pre-specified targeted HRQoL dimensions were GHS (mITT1), physical (mITT2) and emotional functioning (mITT3), fatigue (mITT4) and pain (mITT5).

All tests were two-sided and the type I error was set to 0.05 except for the HRQoL analysis integrating the 5 pre-specified targeted HRQoL dimensions. Regarding these analyses, the type I error was set to 0.01 in order to prevent inflation of alpha risk due to multiple comparisons (Bonferroni adjustment, 5 HRQoL dimensions). A five-point difference in HRQoL scores was considered as the Minimal Clinically Important Difference (MCID) [25].

Descriptive analysis

Baseline variables were described using means and standard deviations (SD) for continuous variables and percentages for qualitative variables. Baseline HRQoL scores were described by treatment arm and compared with a Mann-Whitney non-parametric test. The number of HRQoL questionnaires completed at each measurement time was reported. The Most Common Grade 3 or 4 Adverse Events occurring during the study according to the National Cancer Institute Common Terminology Criteria for Adverse Events (version 3.0) [26] were reported by treatment arm.

Missing data analysis

The MD patterns were patients with at least one missing HRQoL score during the follow-up (partial responders) versus patients with all available scores until their drop-out of the study or death (complete responders). The number and percentage of patients according to the MD profile (partial vs. complete responders) were described at each measurement time by treatment arm. The number and percentage of complete responders, partial responders and non responders (patients who did not complete any HRQoL questionnaire) were described by treatment arm and the difference between the two treatment arms was compared using Chi-square test. All baseline variables that could be associated with MD patterns (partial vs. complete responders) were tested with an univariate logistic regression model. Variables with an univariate *P*-value ≤ 0.20 were eligible for multivariate analysis. To prevent collinearity, when two variables were significantly correlated, one variable was retained according to its clinical relevance. The final multivariate model was chosen according to the Akaike criteria and the area under the ROC curve and described with Odds-Ratio (OR) and its 95% confidence interval (CI). Fitted values were then extracted from the model and constituted the PS [27].

Longitudinal analysis

The QFS was defined as the time from randomization to a first deterioration with a 5-point MCID as compared to the baseline score with no further improvement of more than 5 points as compared to the baseline sore, or all-cause of death [8]. Patients with no baseline score were censored at baseline (Day 0). Patients with no follow-up measure were censored just after baseline (Day 1). Patients with no deterioration before their drop-out and those with a deterioration followed by a significant improvement are censored at the time of the last follow-up or the last HRQoL assessment. Each targeted dimensions of the QLQ-C30 was studied.

Based on the intention-to treat principle and according to the worst possible scenario, a sensitivity analysis was performed integrating non responders patients and considering these patients in deterioration since baseline (Day 1).

QFS curves were calculated using the Kaplan-Meier estimation method and described using median and its 95% CI. QFS were compared according to treatment arm using the log-rank test and univariate Cox analyses to estimate HR with 95% CI. Follow-up was calculated using reverse Kaplan-Meier estimation.

To take into account MD, analyses were repeated by assigning a weight to patients according to the IPW method of PS [21]. The weight equals to the inverse of the PS value for partial responders and to the inverse of the opposite of the PS value for complete responders [21].

Multivariate Cox regression model was conducted in order to investigate the independent prognostic value of treatment arm of QFS. All variables collected at baseline were tested in univariate analysis. Some interaction effects between treatment arm and clinical variables were investigated. Variables with an univariate *P*-value ≤ 0.20 were eligible for multivariate analysis. The same variables were kept in multivariate analysis for unweighted and weighted QFS analyses. The variable treatment arm was forced in multivariate analysis.

All analyses were performed with R software [28].

RESULTS

Study population

Between October 2007 and May 2011, 98 patients (49 in each treatment arm) were enrolled in 10 French centers (Fig.1). Baseline characteristics of patients are summarized in Table 1. The median age was 62 years (38-76) and 59 patients (60.20%) were men. At baseline, 34 patients (69.4%) completed the QLQ-C30 questionnaire in Arm 1 and 30 patients (61.2%) in Arm 2. No difference of baseline HRQoL level was observed at baseline between treatment arms (Table S1 in File S1).

The median follow up was 32.5 months (95%CI 25.4-40.4).

The primary endpoint (6-month PFS rate), on 46 first randomized patients per arm using Fleming's criterion was reached in Arm 2 with 20 patients alive and free of progression resulting in an observed 6-month PFS rate of 43.5% [95%CI 12.9-39.3] but not in Arm 1 with only 12 patients alive and free of progression resulting in a 6-month PFS rate of 26.1% [12.9 % - 39.3 %].

Among all randomized patients, the estimated 6 months PFS rate was 25.7% [14.4-38.6] for Arm 1 and 44.9% [30.7-58.0] for Arm 2. The objective response rate was 10.2% [1.4-19.0] for Arm 1 and 36.7% [22.7-50.7] for Arm 2. Median OS was 8.2 months [95%CI 5.3 -9.2] in Arm 1 and 11 months [7.8-13.6] in Arm 2 [1].

The Most Common Grade 3 or 4 Adverse Events occurring during the study are reported in Table S2 in File S1.

Missing data analysis

Table 2 gives the number and percentage of complete, partial and non-responders in each treatment arm. The difference of proportion between the two treatment arms was not statistically significant (*P*-value=0.12 Chi-square test).

Among the 66 patients (67.3%) who had completed at least one HRQoL questionnaire during the study, 40 (60.6%) were partial responders (15 in Arm 1 (37.5%), 25 in Arm 2 (62.5%)) and 26 (39.4%) were complete responders (15 in Arm 1 (57.7%), 11 in Arm 2 (42.3%)) during the follow-up. The details of the HRQoL questionnaire completed at each follow-up measurement time according to treatment arm and MD profile are given in Table 3.

Based on the univariate analyses, variables associated with responder profiles and retained to build the PS were a primary tumor location at the pancreatic head (yes vs. no), presence of metastatic lymph node (yes vs. no), neutrophils, hemoglobin and platelet rates (dichotomized according to the median value). In multivariate analysis, a primary tumor location at the pancreatic head (OR = 2.72 [0.86-9.16]), the presence of lymph node metastases (7.90 [1.12-164.12]), a low neutrophils (2.13 [0.64-7.25]) and platelets rate (2.77 [0.84-9.72]) and a high hemoglobin rate (1.80

[0.54-6.10]) were independently associated with partial responder profile but not statistically significant. The area under the ROC curve was equal to 0.76.

Longitudinal analysis

In Arm 1 (gemcitabine alone) and Arm 2 (gemcitabine + FOLFIRI.3) respectively:

- 18 and 17 patients experienced a QFS of GHS, among the 63 patients retained (30 in Arm 1, 33 in Arm 2).

- 19 and 17 patients experienced a QFS of physical functioning, among the 65 patients retained (30 in Arm 1, 35 in Arm 2)

- 19 and 18 patients experienced a QFS of emotional functioning, among the 63 patients retained (30 in Arm 1, 33 in Arm 2).

- 18 and 17 patients experienced a QFS of fatigue, among the 65 patients retained (30 in Arm 1, 35 in Arm 2).

- 18 and 16 patients experienced a QFS of pain, among the 65 patients retained (30 in Arm 1, 35 in Arm 2).

Regarding the unweighted analysis (Table 4), patients in FOLFIRI.3 + gemcitabine regimen presented a significantly longer QFS than those of Arm 1 only for physical functioning (HR = 0.40 [CI 95% 0.20-0.82], *P*-value<0.01) with a median QFS of 7.92 months [4.21-13.6] for Arm 1 and 9.42 [3.81-13.47] for Arm 2. Regarding the weighted analysis, the same result was observed and patients in Arm 2 presented significantly longer QFS of GHS (HR = 0.52 [0.31-0.85]), emotional functioning (HR = 0.35 [0.21-0.59]), and pain (HR = 0.50 [0.31-0.81]). The median QFS of GHS was 4.34 months [4.21-9.72] for Arm 1 and 12.06 [9.46-13.47] for Arm 2. The median QFS of emotional functioning was 4.27 months [4.04-7.92] for Arm 1 and 12.48 [9.46-22.57] for Arm 2. Regarding pain, the median QFS was 7.92 months [4.21-9.49] for Arm 1 and 11.60 [9.46-13.21] for Arm 2. These QFS curves were described in Figure 2.

Variables retained for the Cox multivariate analysis were treatment arm (Arm 2 vs. Arm 1), number of metastatic sites (2 or more vs. 1) and an interaction effect

between treatment arm and the number of metastatic sites, according to the univariate Cox regression analysis (data not shown).

No results were statistically significant in the unweighted analysis. Regarding the weighted analyses, the treatment arm (gemcitabine + FOLFIRI.3) and the number of metastatic sites (one site) were independently associated with longer QFS of physical functioning (Table 5). The number of metastatic sites (more than one vs. one) remained significantly associated with a shorter QFS of GHS, fatigue and pain.

As for the unweighted analysis, the same trend were observed for the sensitivity unweighted analysis integrating non-responders patients but not statistically significant (see Figure S1, Table S3 and S4 in File S1).

DISCUSSION

As previously reported, patients treated with sequential chemotherapy FOLFIRI.3 + gemcitabine presented a benefit in PFS at 6 months (44.9% (30.7-58.0) vs. 25.7% (14.4-38.6)), OS (64.7%(49.5-76.4) vs. 62.8% (47.6-74.7)) and objective response rate (36.7% vs. 10.2%) [1].

Meanwhile to the recent progress in the improvement of OS, preserving HRQoL is of paramount importance considering the symptom burden and the poor prognosis of mPC. If several Phase III trials attempted to show a clinical benefit or improvement in HRQoL, few have achieved their goals [29,30]. Recently, the clinical trial comparing FOLFIRINOX to gemcitabine shown an improvement in HRQoL for FOLFIRINOX arm [5].

In our trial, HRQoL results support the efficacy profile of FOLFIRI.3 + Gemcitabine regimen. Patients in FOLFIRI.3 + Gemcitabine arm presented a longer QFS than those of gemcitabine alone arm whatever the HRQoL score considered in both QFS analyses. In multivariate weighted analysis, treatment with sequential FOLFIRI.3 + gemcitabine was significantly associated with longer QFS in each HRQoL score considered including pain and fatigue score, two symptoms commonly present at time of diagnosis.

Median QFS for each domain was shorter than median PFS irrespective of the use of IPW method. It is noteworthy that survival estimates depend on the QFS definition. Contrary to our definition, all-cause death was not integrated as an event in the definition of TUDD chosen by Gourgou-Bourgade et al. [5]. In consequence, median TUDD was not reached after a 26.6 months follow-up while the median PFS was 6.4 months in the FOLFIRINOX arm [2,5] which was not in agreement with clinical profiles of these patients. Moreover it is underlined that comparison across trials is not possible, stressing the need to adopt a common definition of TUDD or QFS [9].

If QFS is increasingly used in clinical trials, consensual methods to optimize management of MD are still lacking [5,31,32,33]. In FOLFIRINOX trial, little information was provided on the method used to deal with MD, except when authors declared that the two groups did not differ in terms of rate of MD [5].

In our study, in both unweighted and weighted analyses, patients in Arm 2 presented a longer QFS than patients in Arm 1. In multivariate analyses, treatment arm (gemcitabine + FOLFIRI.3) and number of metastatic sites (one site) were significantly associated with longer QFS of physical functioning in the weighted analysis. The same trend were observed for the unweighted analysis but not statistically significant.

In this way, using the IPW method of the PS influences the results of the multivariate analysis by underlining more significant associations. A high weight is assigned to patients with no MD (mainly patients of Arm 1) and a low weight to partial responders (mainly patients of Arm 2). As in unweighted analysis, a longer QFS was yet observed for patients of Arm 2 as compared to those of Arm 1 for most HRQoL dimensions, the HR increased with the use of the IPW method. The use of the PS in conjunction with the TUDD method allowed reducing the bias due to the occurrence of MD depending on patients' characteristics during the follow-up. This bias cannot be totally eliminated because MD can also depend on unobserved data. However, some logistic problems could explain the reasons for partial and non-responders because these patients were followed in the study for other endpoints.

Besides primary prevention procedures for limiting MD rate, additional work on statistical methods to handle with MD is still needed. Multiple imputations on the HRQoL scores could also be performed but this method requires a larger sample and

can only retain one or two factors associated with MD [34], more variable can be retained in the PS. Then this approach could be suggested for the trials with limited sample size. Contrary to the pattern mixture models, the IPW method in conjunction with the TUDD approach is more appropriate to the design of oncology clinical trials, for which a lot of HRQoL measures are done. In fact, the number of possible patterns increases with the number of HRQoL measures. Austin et al. recommend to use IPW for time to event data [21]. Propensity score matching could also be performed for survival analysis but a higher sample size is needed. Finally, the IPW method is easy understandable (weighting observations according to the presence or absence of MD) [21].

In conclusion, analyses of QFS supports that sequential strategy with FOLFIRI.3 followed by gemcitabine in patients with untreated mPC is feasible and, despite more toxicities, delayed the HRQoL deterioration. Moreover, using the PS allows controlling the imbalance of informative MD between the two arms and provides more precise estimation of the treatment effect. This sequential treatment strategy will now be compared with FOLFIRINOX in a phase III trial (French study).

Acknowledgements

Financial support for this research was provided by Pfizer, AGEO and AROLD (Association pour la Recherche en Oncologie Digestive). This study was also supported by a grant from the French Public Health Research Institute (http://www.iresp.net) under the 2012 call for projects as part of the 2009-2013 Cancer Plan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. Trouilloud I, Dupont-Gossard AC , Artru P, Lecomte T, Gauthier M, Aparicio T, Thirot-Bidault A, Malka D, Lobry C, Asnacios A, Lacombe S, Fein F, Fanica D, Dubreuil O, Marthey L, Zaanan A, Bonnetain F, Taïeb J (2012) FOLFIRI.3 alternating with gemcitabine or gemcitabine alone in patients with previously untreated metastatic pancreatic adenocarcinoma: Results of the randomized multicenter AGEO phase II trial FIRGEM. J Clin Oncol 30, 2012 (suppl; abstr 4018).
2. Conroy T, Desseigne F, Ychou M, Bouche O, Guimbaud R, et al. (2011) FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. N Engl J Med 364: 1817-1825.
3. Von Hoff DD, Ervin T, Arena FP, Chiorean EG, Infante J, et al. (2013) Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. New England Journal of Medicine.
4. Seufferlein T, Bachet JB, Van Cutsem E, Rougier P, Group EGW (2012) Pancreatic adenocarcinoma: ESMO-ESDO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 23 Suppl 7: vii33-40.
5. Gourgou-Bourgade S, Bascoul-Mollevi C, Desseigne F, Ychou M, Bouche O, et al. (2013) Impact of FOLFIRINOX compared with gemcitabine on quality of life in patients with metastatic pancreatic cancer: results from the PRODIGE 4/ACCORD 11 randomized trial. J Clin Oncol 31: 23-29.
6. Fiteni F, Westeel V, Pivot X, Borg C, Vernerey D, et al. (2014) Endpoints in cancer clinical trials. J Visc Surg 151: 17-22.
7. Bonnetain F, Bosset JF, Gerard JP, Calais G, Conroy T, et al. (2012) What is the clinical benefit of preoperative chemoradiotherapy with 5FU/leucovorin for T3-4 rectal cancer in a pooled analysis of EORTC 22921 and FFCD 9203 trials: surrogacy in question? Eur J Cancer 48: 1781-1790.
8. Bonnetain F, Dahan L, Maillard E, Ychou M, Mitry E, et al. (2010) Time until definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. Eur J Cancer 46: 2753-2762.
9. Anota A, Hamidou Z, Paget-Bailly S, Chibaudel B, Bascoul-Mollevi C, et al. (2013) Time to health-related quality of life score deterioration as a modality of

longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization? Quality of Life Research.

10. Little RJ, Rubin DB (1987) Statistical analysis with missing data. New York: John Wiley & Sons.
11. Sherman DW, McSherry CB, Parkas V, Ye XY, Calabrese M, et al. (2005) Recruitment and retention in a longitudinal palliative care study. Appl Nurs Res 18: 167-177.
12. Fairclough DL, Peterson HF, Chang V (1998) Why are missing quality of life data a problem in clinical trials of cancer therapy? Stat Med 17: 667-677.
13. Ross L, Thomsen BL, Boesen EH, Johansen C (2004) In a randomized controlled trial, missing data led to biased results regarding anxiety. J Clin Epidemiol 57: 1131-1137.
14. Curran D, Bacchi M, Schmitz SF, Molenberghs G, Sylvester RJ (1998) Identifying the types of missingness in quality of life data from clinical trials. Stat Med 17: 739-756.
15. Liao K, Freres DR, Troxel AB (2012) A transition model for quality-of-life data with non-ignorable non-monotone missing data. Stat Med 31: 3444-3466.
16. Fairclough DL, Peterson HF, Cella D, Bonomi P (1998) Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. Stat Med 17: 781-796.
17. Pauler DK, McCoy S, Moinpour C (2003) Pattern mixture models for longitudinal quality of life studies in advanced stage disease. Stat Med 22: 795-809.
18. Little RJ, Wang Y (1996) Pattern-mixture models for multivariate incomplete data with covariates. Biometrics 52: 98-111.
19. Trojano M, Pellegrini F, Paolicelli D, Fuiani A, Di Renzo V (2009) observational studies: propensity score analysis of non-randomized data. Int MS J 16: 90-97.
20. D'Agostino RB, Jr. (1998) Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 17: 2265-2281.
21. Austin PC (2013) The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med 32: 2837-2849.
22. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, et al. (1993) The European Organization for Research and Treatment of Cancer QLQ-C30: a

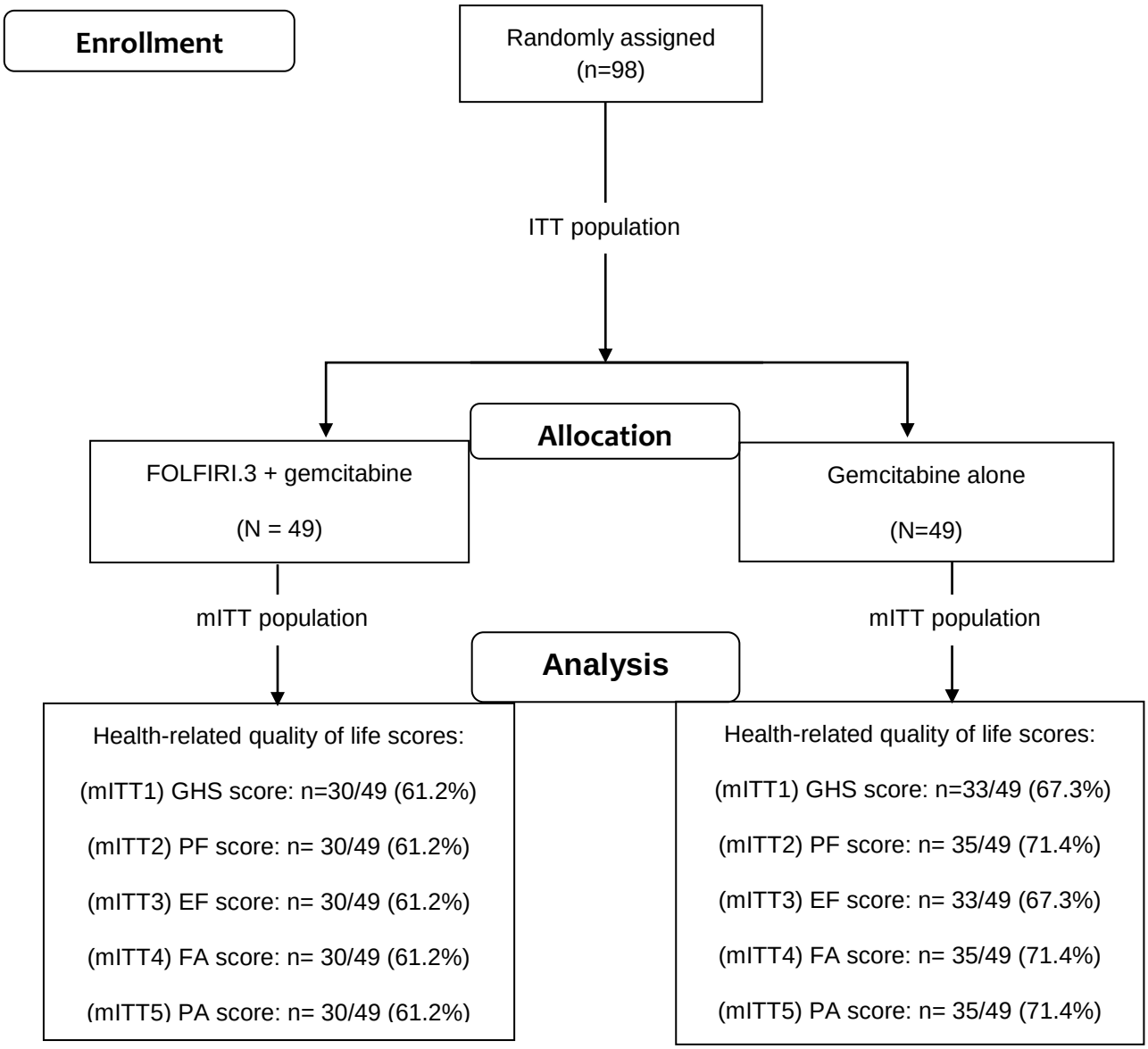
quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 85: 365-376.

23. Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, et al. (2001) EORTC QLQ-C30 Scoring Manual (3rd edition). Brussels: EORTC 2001 ed2001.
24. Fleming TR (1982) One-sample multiple testing procedure for phase II clinical trials. Biometrics 38: 143-151.
25. Osoba D, Rodrigues G, Myles J, Zee B, Pater J (1998) Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 16: 139-144.
26. NCI (2006) Common Terminology Criteria for Adverse Events v3.0.
27. Little RJ, Rubin DB (2000) Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annual review of public health 21: 121-145.
28. Team RDC (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
29. Moinpour CM, Vaught NL, Goldman B, Redman MW, Philip PA, et al. (2010) Pain and emotional well-being outcomes in Southwest Oncology Group-directed intergroup trial S0205: a phase III study comparing gemcitabine plus cetuximab versus gemcitabine as first-line therapy in patients with advanced pancreas cancer. J Clin Oncol 28: 3611-3616.
30. Bernhard J, Dietrich D, Scheithauer W, Gerber D, Bodoky G, et al. (2008) Clinical benefit and quality of life in patients with advanced pancreatic cancer receiving gemcitabine plus capecitabine versus gemcitabine alone: a randomized multicenter phase III clinical trial--SAKK 44/00-CECOG/PAN.1.3.001. J Clin Oncol 26: 3695-3701.
31. Cortes J, Baselga J, Im YH, Im SA, Pivot X, et al. (2013) Health-related quality-of-life assessment in CLEOPATRA, a phase III study combining pertuzumab with trastuzumab and docetaxel in metastatic breast cancer. Ann Oncol.
32. Yang JC, Hirsh V, Schuler M, Yamamoto N, O'Byrne KJ, et al. (2013) Symptom Control and Quality of Life in LUX-Lung 3: A Phase III Study of Afatinib or Cisplatin/Pemetrexed in Patients With Advanced Lung Adenocarcinoma With EGFR Mutations. J Clin Oncol.

33. Burris HA, 3rd, Lebrun F, Rugo HS, Beck JT, Piccart M, et al. (2013) Health-related quality of life of patients with advanced breast cancer treated with everolimus plus exemestane versus placebo plus exemestane in the phase 3, randomized, controlled, BOLERO-2 trial. Cancer 119: 1908-1915.
34. Fairclough DL (2010) Design and analysis of quality of life studies in clinical trials: CRC press.

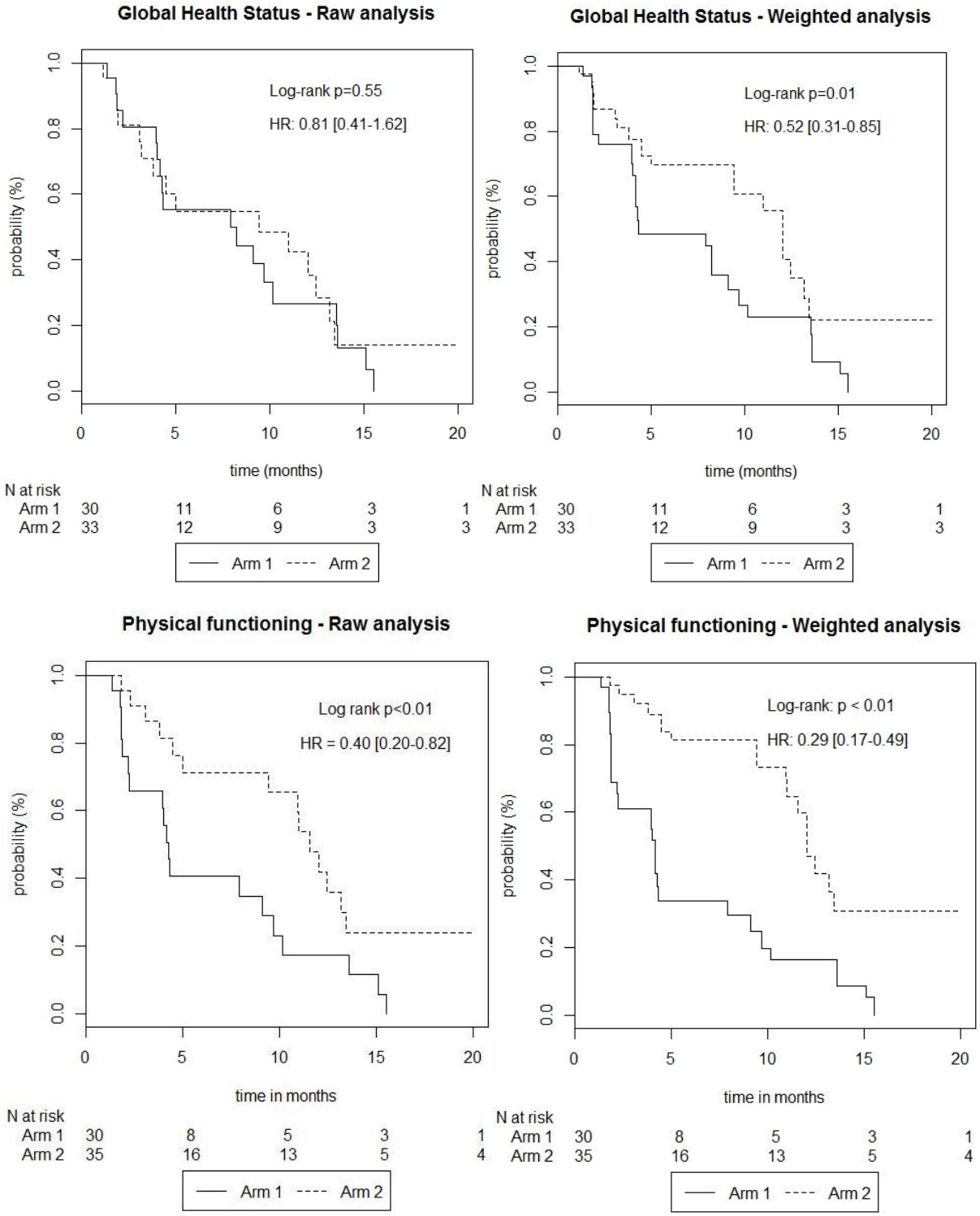
Figure legends

Figure 1. CONSORT Diagram for health-related quality of life analysis.

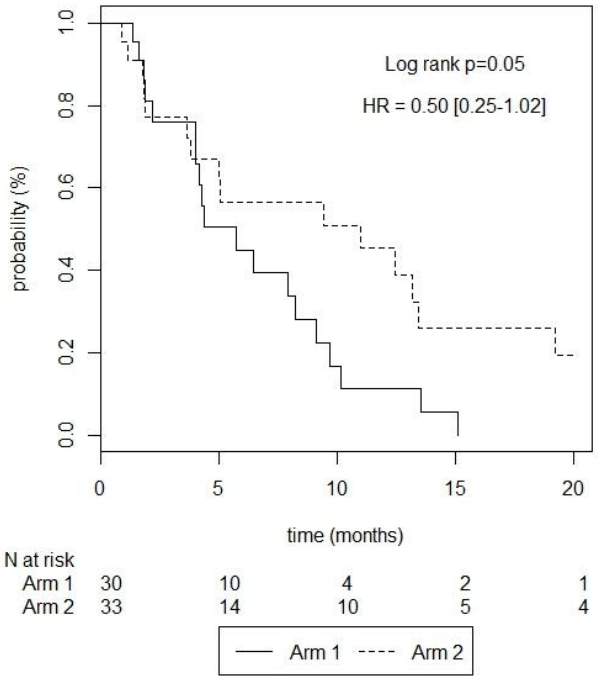


ITT: intent to treat; mITT modified intent to treat; GHS global health status; PF: Physical functioning; EF: Emotional functioning; FA: Fatigue; PA: Pain

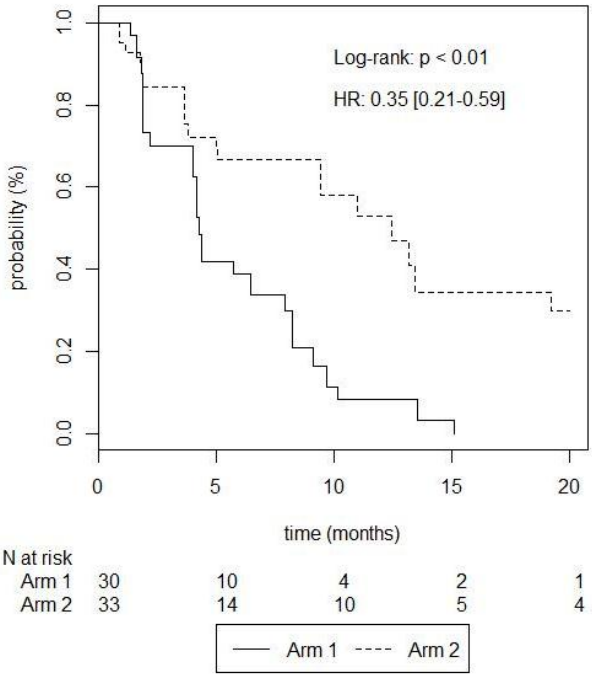
Figure 2. Kaplan-Meier survival curves of the HRQoL deterioration-free survival by treatment arm[a] for the raw and the weighted analysis.



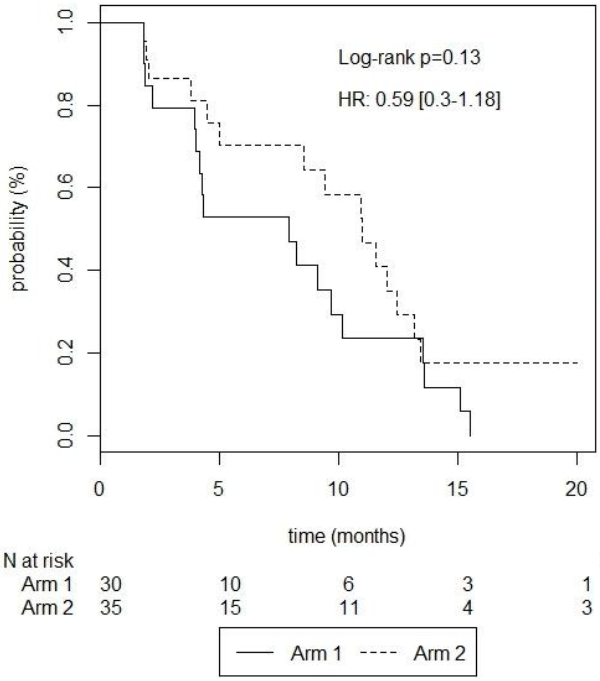
Emotional functioning - Raw analysis



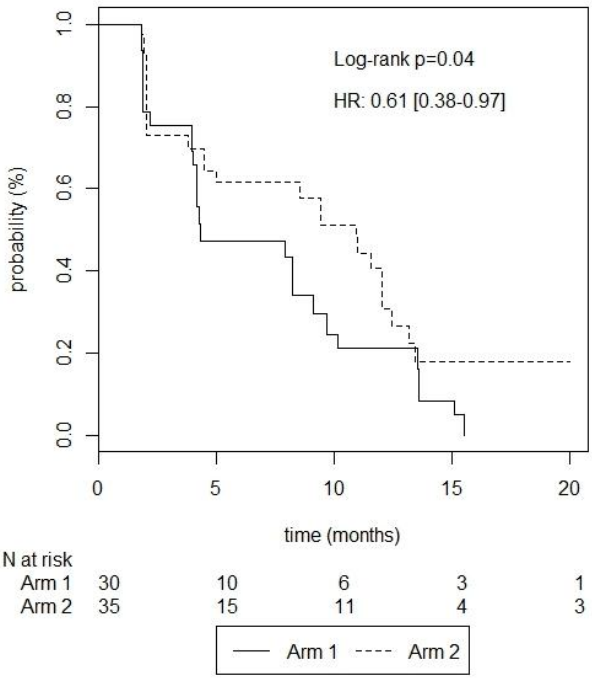
Emotional functioning - Weighted analysis

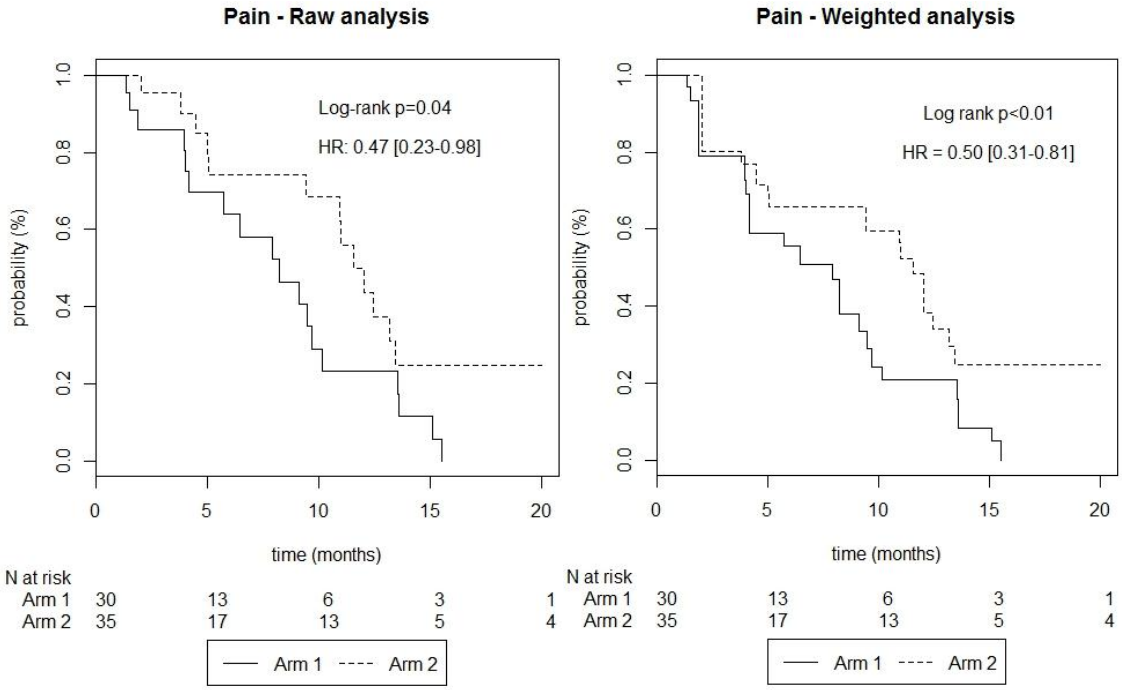


Fatigue - Raw analysis



Fatigue - Weighted analysis





[a]Arm 1: gemcitabine alone, Arm 2: gemcitabine + FOLFIRI.3

Tables

Table 1: Baseline characteristics of patients included according to treatment arm

Variable	Response Category	Arm1 gemcitabine alone (N=49)		Arm 2 FOLFIRI.3 + gemcitabine (N=49)	
		N	%	*N*	%
Sex	male	28	57.1	31	63.3
	female	21	42.9	18	36.7
WHO performance status	0	16	32.6	16	32.7
	1	33	67.4	33	67.3
Previous surgery	no	40	81.6	38	77.6
	yes	9	18.4	11	22.4
Surgery type	curative	4	8.2	5	10.2
	palliative	5	10.2	5	10.2
	not applicable	40	81.6	38	77.6
	missing	0	0.0	1	2.0
Number of metastatic sites	1	35	71.4	33	67.3
	more than 1	14	28.6	16	32.7
Previous chemotherapy	no	40	81.6	44	89.8
	yes	3	6.1	2	4.1
	missing	6	12.2	3	6.1
Primary tumor location	head	29	59.2	18	36.7
	body	11	22.5	17	34.7
	tail	12	24.5	17	34.7
Sites of metastasis	liver	35	71.4	39	79.6
	lung	11	22.5	11	22.5
	lymph node	7	14.3	5	10.2
	peritoneal	10	20.4	16	33.7
	other	2	4.0	3	6.1
Age (years)[a]		45	63 [41-76]	48	62 [38-76]
Leukocytes (/mm3) [a]		49	7600 [3100-36500]	49	8300 [85-21700]
neutrophils (/mm3) [a]		49	5000 [1800-32850]	48	5591.5 [2300-19530]
Creatinine (µmol/l) [a]		48	71.0 [39-105]	48	70 [45-108]
Glycaemia (mmol/l) [a]		29	6.2 [4.1-14]	25	5.8 [0.7-15]
Bilirubin (µmol/l) [a]		48	11.6 [4-227]	45	12 [1-154]
LDH (UI/L) [a]		23	271 [96-5022]	22	340.5 [133-766]
Hemoglobin (g/dl) [a]		49	12.8 [7.9-16.5]	48	12.9 [9.4-16]
Platelet (10^3/mm3) [a]		49	239 [94-570]	48	278.5 [111-634]
ASAT (UI/L) [a]		49	26 [8-149]	46	41.5 [10-187]
ALAT(UI/L) [a]		49	35 [8-155]	46	53.5 [10-348]
Prothrombin (%)[a]		39	93 [26-109]	41	86 [19-122]

[a] **Median [min-max] for continuous variables**

Table 2. Proportion of complete, partial and non responders for HRQoL assessment in each treatment arm

	complete responders (N=26)	partial responders (N=40)	non responders (N=32)
	N (%)	N (%)	N (%)
Arm gemcitabine alone	15 (57.7)	15 (37.5)	19 (59.4)
Arm FOLFIRI + gemcitabine	11 (42.3)	25 (62.5)	13 (40.6)

Table 3. Completion of HRQoL questionnaire at each follow-up measurement time according to treatment arm and missing data profile

	Complete responders			Partial responders		
	Arm 1	Arm 2	Total	Arm 1	Arm 2	Total
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
cycle 1	14 (44.1)	15 (50.0)	29 (46.0)	19 (55.9)	15 (50.0)	34 (54.0)
cycle 2	13 (46.4)	7 (35.0)	20 (41.2)	15 (53.6)	13 (65.0)	28 (58.3)
cycle 3	7 (35.0)	1 (10.0)	8 (26.7)	13 (65.0)	9 (90.0)	22 (73.3)
cycle 4	4 (28.6)	1 (50.0)	5 (31.3)	10 (71.4)	1 (50.0)	11 (68.8)
cycle 5	3 (50.0)	1 (50.0)	4 (50.0)	3 (50.0)	1 (50.0)	4 (50.0)
cycle 6	1 (33.3)	0 (0.0)	1 (25.0)	2 (66.7)	1 (100.0)	3 (75.0)
cycle 7	1 (50.0)	0 (0.0)	1 (50.0)	1 (50.0)	0 (0.0)	1 (50.0)
cycle 8	0 (0.0)	0 (0.0)	0 (0.0)	1 (100.0)	0 (0.0)	1 (100.0)
cycle 9	1 (100.0)	0 (0.0)	1 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
cycle 10	1 (100.0)	0 (0.0)	1 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)

Arm 1: gemcitabine alone; Arm 2: gemcitabine + FOLFIRI.3

Table 4: Results of the Kaplan-Meier estimation of the health-related quality of life deterioration-free survival for a QLQ-C30 score and comparison between treatment arms

		unweighted analysis				weighted analysis		
		N (events)	median [CI 95%]	Log-rank	HR [CI 95%]	median [CI 95%]	Log-rank	HR [CI 95%]
Global health status	Arm 1 [a]	30 (18)	7.92 |4.21-13.6]	0.55	1	4.34 [4.21-9.72]	<0.01	1
	Arm 2 [b]	33 (17)	9.46 |3.81-13.47]		0.81 (0.41-1.62)	12.06 [9.46-13.47]		0.52 [0.31-0.85]
Physical functioning	Arm 1	30 (19)	4.27 [2.27-10.15]	<0.01	1	4.21 [2.27-7.92]	<0.01	1
	Arm 2	35 (17)	11.6 [9.46-26.25]		0.40 (0.2-0.82)	12.06 [11.6-22.57]		0.29 [0.17-0.49]
Emotional functioning	Arm 1	30 (19)	5.75 [4.04-9.72]	0.05	1	4.27 [4.04-7.92]	<0.01	1
	Arm 2	33 (18)	11.01 [3.81-22.57]		0.50 (0.25-1.02)	12.48 [9.46-22.57]		0.35 [0.21-0.59]
Fatigue	Arm 1	30 (18)	7.92 [4.21-13.57]	0.13	1	4.34 [4.21-9.13]	0.04	1
	Arm 2	35 (17)	11.01 [8.57-13.47]		0.59 (0.30-1.18)	10.97 [5.03-12.06]		0.61 [0.38-0.97]
Pain	Arm 1	30 (18)	8.25 [5.75-13.57]	0.04	1	7.92 [4.21-9.49]	<0.01	1
	Arm 2	35 (16)	11.6 [10.97-NA]		0.47 (0.23-0.98)	11.6 [9.46-13.21]		0.50 |0.31-0.81]

[a] Arm 1: gemcitabine alone;

[b] Arm 2: gemcitabine+FOLRIRI.3

Significance level P-value <0.01 according to Bonferroni adjustment

Table 5: Results of the multivariate Cox regression analysis for the QFS analysis of each targeted score of the QLQ-C30 for the raw and the weighed analysis

			without IPW		with IPW	
		***N* (events)**	**HR [CI 95%]**	***P***	**HR [CI 95%]**	***P***
Global health status		63 (35)				
arm [a]	(arm 2) vs.(arm 1)		0.86 [0.38 - 1.96]	0.73	0.58 [0.31 -1.07]	0.08
number of metastatic sites	(2 or more) vs. 1		3.98 [1.21 - 13.71]	0.02	4.39 [2.03 - 9.49]	**<0.01**
Interaction between arm and number of metastatic sites			0.38 [0.08 - 1.88]	0.23	0.41 [0.13 - 1.27]	0.12
Physical functioning		65 (36)				
arm [a]	(arm 2) vs.(arm 1)		0.34 [0.14 – 0.82]	0.02	0.25 [0.13 - 0.48]	**<0.01**
number of metastatic sites	(2 or more) vs. 1		2.80 [0.87 - 9.08]	0.09	2.71 [1.28 - 5.75]	**<0.01**
Interaction between arm and number of metastatic sites			0.86 [0.18 - 4.16]	0.85	1.09 [0.36 - 3.30]	0.87
Emotional functioning		63 (37)				
arm [a]	(arm 2) vs.(arm 1)		0.44 [0.19 – 1.03]	0.06	0.29 [0.15 - 0.56]	**<0.01**
number of metastatic sites	(2 or more) vs. 1		2.72 [0.84 - 8.79]	0.09	2.59 [1.24 - 5.41]	0.01
Interaction between arm and number of metastatic sites			0.91 [0.19 - 4.47]	0.91	1.47 [0.50 - 4.37]	0.49
Fatigue		65 (35)				
arm [a]	(arm 2) vs.(arm 1)		0.54 [0.23 - 1.24]	0.14	0.71 [0.40 - 1.24]	0.22
number of metastatic sites	(2 or more) vs. 1		3.27 [0.86 - 12.42]	0.08	3.40 [1.58 - 7.30]	**<0.01**
Interaction between arm and number of metastatic sites			0.58 [0.11 - 3.17]	0.53	0.39 [0.13 - 1.11]	0.08
Pain		65 (34)				
arm [a]	(arm 2) vs.(arm 1)		0.44 [0.18 - 1.07]	0.07	0.57 [0.32 - 1.03]	0.06
number of metastatic sites	(2 or more) vs. 1		3.04 [0.93 - 9.91]	0.07	3.15 [1.51 - 6.57]	**<0.01**
Interaction between arm and number of metastatic sites			0.66 [0.13 - 3.31]	0.61	0.46 [0.16 - 1.33]	0.15

[a] Arm 1: gemcitabine alone, Arm 2: gemcitabine + FOLFIRI.3

Significance level P-value <0.01 according to Bonferroni adjustment

Supplementary file

Table S1: health-related quality of life scores at baseline according to treatment arm

	N	Mean (SD)	P-value Mann-Whitney
Global Health Status			
Arm gemcitabine alone	28	59.2 (23.3)	0.75
Arm FOLFIRI + gemcitabine	29	56.3 (21.7)	
Physical Functioning			
Arm gemcitabine alone	28	75.4 (23.6)	0.54
Arm FOLFIRI + gemcitabine	31	77.9 (22.4)	
Emotional Functioning			
Arm gemcitabine alone	28	65.3 (23.3)	0.89
Arm FOLFIRI + gemcitabine	29	66.3 (20.7)	
Fatigue			
Arm gemcitabine alone	28	49.1 (28.9)	0.72
Arm FOLFIRI + gemcitabine	31	45.6 (31.8)	
Pain			
Arm gemcitabine alone	28	42.9 (32.2)	0.46
Arm FOLFIRI + gemcitabine	31	36.6 (25.3)	

Table S2: Most Common Grade 3 or 4 Adverse Events according to treatment arm

	Arm 1 gemcitabine alone N=49	**Arm 2 FOLFIRI.3 + gemcitabine N=49**
	N (%)	**N (%)**
All toxicities grade >2	32 (65.3)	38 (77.5)
Hematologic toxicity		
Neutropenia	12 (24.5)	24 (49.0)
Febrile neutropenia	0 (0.0)	2 (4.1)
Thrombocytopenia	9 (18.8)	9 (18.4)
Anaemia	3 (6.0)	6 (12.2)
Gastrointestinal toxicity		
Nausea and vomiting	2 (4.1)	4 (8.2)
Diarrhoea	0 (0.0)	6 (12.2)

Table S3: Results of the Kaplan-Meier estimation of the health-related quality of life deterioration-free survival for a QLQ-C30 score considering non-responders patients in deterioration since baseline and comparison between treatment arms

Scores	Treatment arm	*N* (events)	Median [CI 95%]	Log-rank	HR [CI 95%]
Global health status	Arm 1 [a]	49 (37)	2.20 [0.03 - 7.92]	0.54	1
	Arm 2 [b]	49 (33)	3.09 [1.15 - 11.01]		0.86 [0.53-1.39)
Physical functioning	Arm 1	49 (38)	1.84 [0.03- 4.21]	0.03	1
	Arm 2	49 (31)	5.03 [2.33 - 12.48]		0.57 [0.35 - 0.95)
Emotional functioning	Arm 1	49 (38)	1.91 [0.03 - 4.37]	0.13	1
	Arm 2	49 (34)	1.91 [0.92 - 11.01]		0.68 [0.42 - 1.12]
Fatigue	Arm 1	49 (37)	2.20 [0.03 - 7.92]	0.15	1
	Arm 2	49 (31)	5.03 [1.97 - 12.06]		0.70 [0.43 - 1.14]
Pain	Arm 1	49 (37)	3.98 [0.03 - 8.25]	0.07	1
	Arm 2	49 (30)	9.46 [3.81 - 12.48]		0.63 [0.38 - 1.04]

[a] Arm 1: gemcitabine alone;

[b] Arm 2: gemcitabine+FOLRIRI.3

Significance level P-value <0.01 according to Bonferroni adjustment

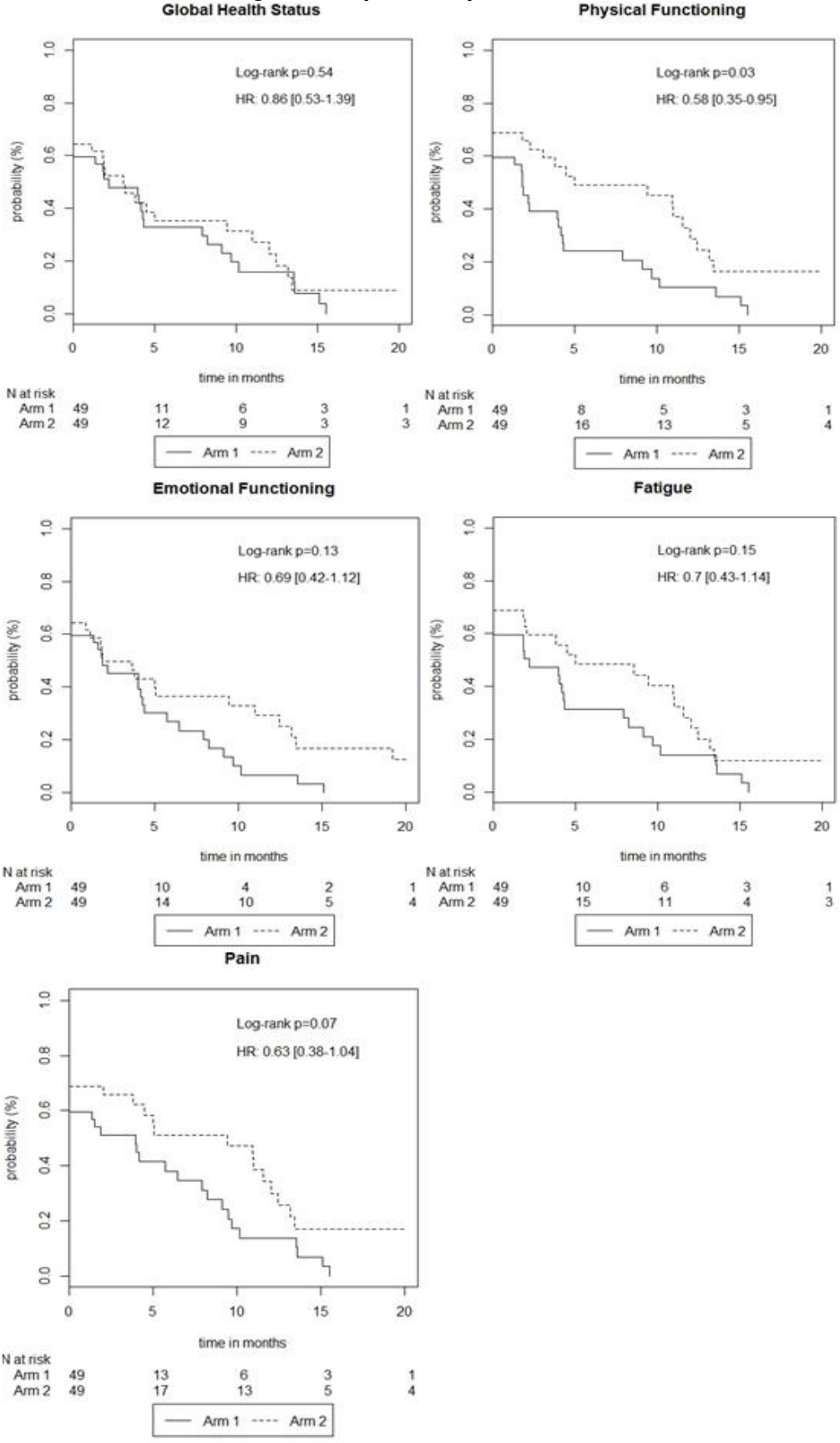
Table S4. Results of the multivariate Cox regression analysis for the QFS analysis of each targeted score of the QLQ-C30 considering non-responders patients in deterioration since baseline

		***N* (events)**	**HR [CI 95%]**	***P***
Global health status		98 (70)		
arm [a]	(arm 2) vs.(arm 1)		0.93 [0.52 - 1.67]	0.80
number of metastatic sites	(2 or more) vs. 1		2.14 [1.03 - 4.44]	0.04
Interaction between arm and number of metastatic sites			0.63 [0.22 - 1.77]	0.38
Physical functioning		98 (69)		
arm [a]	(arm 2) vs.(arm 1)		0.54 [0.29 - 1.01]	0.05
number of metastatic sites	(2 or more) vs. 1		1.79 [0.87 - 3.68]	0.11
Interaction between arm and number of metastatic sites			0.96 [0.34 - 2.71]	0.93
Emotional functioning		98 (72)		
arm [a]	(arm 2) vs.(arm 1)		0.65 [0.36 - 1.19]	0.16
number of metastatic sites	(2 or more) vs. 1		1.80 [0.88 - 3.71]	0.11
Interaction between arm and number of metastatic sites			0.96 [0.34 - 2.69]	0.94
Fatigue		98 (68)		
arm [a]	(arm 2) vs.(arm 1)		0.69 [0.38 - 1.25]	0.22
number of metastatic sites	(2 or more) vs. 1		1.87 [0.88 - 3.97]	0.11
Interaction between arm and number of metastatic sites			0.81 [0.28 - 2.33]	0.69
Pain		98 (67)		
arm [a]	(arm 2) vs.(arm 1)		0.63 [0.34 - 1.17]	0.14
number of metastatic sites	(2 or more) vs. 1		1.88 [0.91 - 3.88]	0.09
Interaction between arm and number of metastatic sites			0.81 [0.28 - 2.35]	0.70

[a] Arm 1: gemcitabine alone, Arm 2: gemcitabine + FOLFIRI.3

Significance level P-value <0.01 according to Bonferroni adjustment

Figure S1. Kaplan-Meier survival curves of the HRQoL deterioration-free survival by treatment arm[a] considering non-responders patients in deterioration since baseline



[a]**Arm 1: gemcitabine alone, Arm 2: gemcitabine + FOLFIRI.3**

2.4. TJD et essai de phase I

Résumé

Contexte

Les études de phase I ont pour objectif de déterminer la dose recommandée (dose maximale tolérée, (DMT)) dans les futurs essais. Les toxicités limitant la dose sont généralement utilisées pour définir la DMT. La grille NCI-CTC AE évaluée par le clinicien est alors utilisée en considérant une toxicité de grade 3 ou 4 comme toxicité limitant la dose. Certaines toxicités modérées observées durant une longue période peuvent impactées la QdV des patients. Ces toxicités ne sont pas pourtant pas prises en compte dans l'évaluation de la toxicité limitant la dose selon la grille CI-CTC AE. Dans ce contexte, la QdV pourrait compléter l'évaluation de la toxicité pour détecter des traitements intolérables. L'objectif de cette étude était d'explorer la valeur ajoutée de la QdV dans un essai de phase I pour identifier la dose recommandée dans les essais de phase II.

Matériels et méthodes

Un essai de phase I de chimioembolisation par des microsphères d'embolisation chargées avec de l'idarubicine a été menée chez des patients atteints d'un carcinome hépatocellulaire non résécable. La DMT a été définie comme le palier de dose d'idarubicine pour lequel 20% des patients maximum ont une dose limitante toxique.

La QdV a été évaluée par le questionnaire EORTC QLQ-C30 à l'inclusion, à J15, J30 et J60 après chimioembolisation. Les dimensions de QdV ciblés étaient la QDV/santé globale, le fonctionnement physique, la fatigue et la douleur. Le temps jusqu'à détérioration d'un score de QdV (TJD) a été investigué comme modalité d'analyse longitudinale. Quatre définitions de TJD ont été explorées, intégrant ou non l'occurrence d'une toxicité de grade ¾ selon la grille NCI-CTC AE :

1. le délai entre la date de randomisation et l'observation d'une première détérioration d'au moins 5 points du score de QdV par rapport au score à l'inclusion;

2. le délai entre la date de randomisation et l'observation d'une première détérioration d'au moins 5 points du score de QdV par rapport au score à l'inclusion, intégrant toute cause de décès comme évènement ;

3. le TJD d'au moins un score de QdV parmi les dimensions ciblées, quel que soit le premier évènement considéré ;

4. le TJD d'au moins un score de QdV parmi les dimensions ciblées ou l'occurrence d'au moins une toxicité de grade 3/4 selon la grille NCI-CTC AE, quel que soit le premier évènement considéré.

Une analyse de sensibilité a également été réalisée considérant les patients sans mesure de suivi en détérioration dès l'inclusion. Pour les définitions de TJD 3 et 4, des analyses de Cox univariées ont été réalisées afin d'identifier les facteurs influençant potentiellement le TJD.

Résultats

21 patients ont été inclus entre mars 2010 et mars 2012: 9 patients ont été traités à une dose d'idarubicine de 5 mg, 6 à 10 mg et 6 à 15 mg. La DMT calculée était de 10 mg. Pour toutes les définitions de TJD considérées, les patients inclus à 10 mg présentaient un TJD plus long qu'à 5 mg pour tous les scores de QdV ciblés. A15 mg, les patients présentaient également un TJD plus court du fonctionnement physique et de la douleur qu'à 5 mg que le décès soit pris en compte ou non dans les évènements. Concernant l'analyse de sensibilité considérant l'absence de mesure de suivi comme évènement, les patients inclus à 15 mg d'idarubicine présentaient une détérioration plus tardive de la QdV comparativement aux patients inclus à 5 mg. Les femmes présentaient une détérioration plus tardive de la douleur que les hommes (HR 14.7 [1.31-164.7]).

Conclusion

Les patients inclus au palier de dose retenu 10 mg d'idarubicine semblent présenter une détérioration plus tardive de la QdV comparativement aux patients inclus aux paliers de dose 5 ou 15 mg. Ces résultats sont donc cohérents avec l'évaluation effectuée par le clinicien pour détéerminer la DMT. Ces résultats montrent l'intérêt de l'étude de la QdV dans les essais de phase I. De plus, cela soulève le problème de la construction d'un questionnaire spécifique pour les essais de phase I davantage focalisé sur les toxicités.

<u>Article:</u> An explorative study to assess added value of the health-related quality of life in a phase I trial: Idarubicin-loaded beads for chemoembolization of hepatocellular carcinoma

Article en cours de rédaction

Amélie Anota[1,2], Mathieu Boulin[3,4], Sandrine Dabakuyo[1,5], Patrick Hillon[3,6], Jean Pierre Cercueil[3,7], Anne Minello[3,6], Jean Louis Jouve[3,6], Xavier Paoletti[7], Laurent Bedenne[3,6], Boris Guiu[3,8], Franck Bonnetain[1,2]

[1] Quality of Life in oncology clinical research Platform, France

[2] Methodological and Quality of Life in Oncology Unit, EA 3181, University Hospital of Besançon, France

[3] INSERM U866, University of Burgundy, Dijon, France

[4] Department of Pharmacy, University Hospital, Dijon, France

[5] Methodological and Biostatistics Unit, EA 4184, Centre Georges Francois Leclerc, Dijon, France

[6] Department of Hepatogastroenterology, University Hospital, Dijon, France

[7] Biostatistics department, Institut Curie, Paris, France

[8] Department of Interventional Radiology, University Hospital, Dijon, France

Corresponding author:

Amélie Anota

Quality of Life and Cancer national Platform

Methodological and Quality of Life in Oncology Unit (EA 3181)

University Hospital of Besançon

France

Email: aanota@chu-besancon.fr

Telephone number: +33381218896

Abstract

Purpose

A phase I dose-escalation trial of transarterial chemoembolization (TACE) with idarubicin-loaded beads was performed in cirrhotic patients with hepatocellular carcinoma. MTD was defined as dose level closest to that causing dose limiting toxicity (DLT) in 20% of patients. In this context, the added value of health-related quality of life (HRQoL) of patients to complement the usual toxicity assessment should be questioned. The objective was to investigate Time to HRQoL deterioration (TTD) in this phase I trial.

Methods

HRQoL was evaluated using the EORTC QLQ-C30 at baseline and at days 15, 30 and 60 after TACE. Targeted dimensions were global health status, physical functioning, fatigue and pain. Several definitions of deterioration were investigated integrating all-cause death or not. TTD was first defined as the time from randomization to a first HRQoL score deterioration with a 5-point MCID as compared to the baseline score. Univariate Cox analyses were performed to identify factors influencing TTD.

Results

Between March 2010 and March 2011, 21 patients were included: 9, 6, and 6 patients were treated at idarubicin dose levels of 5-, 10-, and 15-mg, respectively. Calculated MTD of idarubicin was 10 mg. At 10-mg idarubicin dose level, patients presented a longer TTD than at 5-mg for Global Health Status (HR 0.87 [CI95% 0.14-5.36]), physical functioning (HR 0.67 [0.11-4.13]), fatigue (HR 0.77 [0.13-4.71]) and pain (HR 0.52 [0.09-3.16]).

Conclusions

These results are coherent the idarubicin dose level retained as MTD. These results show the importance to study HRQoL in phase I trials. Moreover, it raises the issue of a specific questionnaire for phase I trial which would be more focused on toxicities.

Keywords: Health-related Quality of Life, clinical trial, phase I, oncology, longitudinal analysis, time to deterioration

Abbreviations

CI: confidence interval

DLT: dose limiting toxicity

ECOG PS: Eastern Cooperative Oncology Group performance status

EORTC: European Organisation for Research and Treatment of Cancer

HCC: hepatocellular carcinoma

HR: Hazard Ratio

HRQoL: health-related quality of life

MTD: maximum tolerated dose

MCID: minimal clinically important difference

RP2D: recommended Phase II dose

SD: standard deviation

TACE: transarterial chemoembolization

TTD: time to deterioration

INTRODUCTION

Phase I trials were usually dedicated to identify the maximum-tolerated dose (MTD) and the recommended Phase II dose (RP2D). Dose limiting toxicities (DLT) were used to define primary endpoint of such trials (1). Generally grade 3/4 were targeted for selected kind of toxicities.

Duration of toxicity occurrence and/or late occurrence of toxicity is usually not taken into account in the definition of DLT. Some moderate toxicities observed during a long period could impair patients' health-related quality of life (HRQoL), this is the case for the new biological agent assessed in oncology (2). These moderate toxicities are also not taken into account in the usual definition of DLT based on National Cancer Institute Common Terminology Criteria for Adverse Events (NCI CTCAE) assessment by the clinicians (3). In this way, the usual definition of the DLT may not be adapted to reflect the patient's feeling regarding the tolerability of the treatment received, resulting in an over- or under-estimation of these toxicities (4, 5). As the result, the dose selected could not be the best RP2D.

In this context, the added value of patients' HRQoL to complement the usual NCI CTCAE scale could be a relevant complementary information to detect intolerable treatments. The added value of HRQoL or other patients-reported outcomes to assess safety of treatment toxicities was proposed several times in the last decade (6). Indeed, the evaluation of HRQoL in cancer clinical trials is now fully recommended in order to investigate the clinical benefit of the new therapeutic strategies for the patient (7, 8).

At this time, HRQoL or other patient-reported outcome has been poorly investigated in oncology phase I clinical trials. Indeed, analyses are more descriptive than longitudinal. Moreover, results were not used to identify RP2D (9). One recent study analyzed longitudinal HRQoL in a phase I clinical trials but only with a comparison between two time points: before and after treatment (10).Time to HRQoL score deterioration (11-13) would be an alternative way to analyze these data in a clinical meaningful way.

A phase I clinical trial investigating the recommended dose for transarterial chemoembolization (TACE) with idarubicin-loaded beads with hepatocellular

carcinoma (HCC) has shown that the maximum-tolerated dose (MTD) is at 10-mg idarubicin level (14). The MTD was defined as the dose at which less than 20% of the patients experienced a dose limiting toxicity (DLT). The DLT was determined according to the usual toxicity assessment using the National Cancer Institute Common Terminology Criteria for Adverse Events (NCI-CTCAE v3.0) (15).

The objective of this study was to explore added value of HRQoL to identify RP2D in a phase I trial. Time to HRQOL deterioration approach was used with several definitions of composite endpoints integrating or not the occurrence of a severe toxicity of grade 3/4 according to the NCI-CTCAE criteria made by the clinician.

METHODS

Study design and objectives

This was a phase I, monocentric, open-label, dose-escalation study of TACE with idarubicin-loaded beads. All patients were fully informed of the study and provided signed written informed consent. The protocol was approved by the ethics committees. The design of this study has been extensively described elsewhere (14).

The primary objective of this study was to determine the MTD of idarubicin loaded after a single TACE session

Eligibility criteria

Patients aged 18 years or older with HCC unsuitable for curative treatments were evaluated for the study. Eligibility criteria included: a confirmed diagnosis of HCC according to the European Association for the Study of the Liver criteria (16), Child-Pugh liver function of A to B7 without ascites nor jaundice and Eastern Cooperative Oncology Group performance status (ECOG PS) of 0 to 1.

Treatment consisted of a single TACE session with injection of 2 mL DEBs (DC Bead® 300-500 µm, Biocompatibles, Surrey, UK) loaded with idarubicin (Zavedos®, Pfizer, Paris, France) at one of five following escalating doses: 5-, 10-, 15-, 20-, or

25-mg. The starting dose level of idarubicin was 10 mg. Idarubicin dose escalation followed a likelihood approach continual reassessment method (17, 18).

DLT was defined as any unacceptable toxicity that was possibly, probably, or definitely attributed to treatment. Unacceptable toxicity was defined as any grade 4-5 adverse event from the following categories: allergy/immunology, blood/bone marrow (for neutrophil and platelet count, grade 4 adverse event was not considered a DLT if was reversible within 7 days), cardiac arrhythmia, general cardiac, coagulation, constitutional symptoms, gastrointestinal, haemorrhage/bleeding, infection, metabolic/ laboratory (except for total bilirubin defined as DLT if >5 the upper limit of normal for 5 consecutive days and for creatinine defined as a DLT if grade ≥3 adverse event), musculoskeletal/soft tissue, ocular/visual, pain. For other categories of the NCI-CTCAE v3.0, any grade ≥3 adverse event defined an unacceptable toxicity.

Using this method, the MTD was defined as the dose at which less than 20% of the patients experienced a DLT within the month after TACE. MTD was also the RP2D.

Cohorts of one patient were sequentially enrolled at one dose level on the basis of the DLT observed within the month after TACE of the previous patient.

Health-related quality of life assessment

HRQoL was evaluated using the European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30 questionnaire. The QLQ-C30 is a validated tool to assess HRQoL in cancer (19). The QLQ-C30 includes 30 items and measures five functional scales (physical, role, emotional, cognitive and social functioning), global health status (GHS), financial difficulties and eight symptom scales (fatigue, nausea and vomiting, pain, dyspnea, insomnia, appetite loss, constipation, diarrhea) (19). HRQoL scores were generated according to the EORTC Scoring Manual (20). These scores vary from 0 (worst) to 100 (best) for the GHS and physical functioning, and from 0 (best) to 100 (worst) for fatigue and pain.

HRQoL was evaluated at baseline and at days 15, 30 and 60 after TACE.

Statistical analysis

Population

Patients with at least one HRQoL score were included in the HRQoL analysis (modified intent to treat (mITT) analysis). Pre-specified targeted HRQoL dimensions were GHS (mITT1), physical functioning (mITT2), fatigue (mITT3) and pain (mITT4).

All tests were two-sided and the type I error was set to 0.05. A five-point difference in EORTC HRQoL scores was considered as the minimal clinically important difference (MCID) (21).

Descriptive analysis

Clinical and sociodemographics variables collected at baseline were described with median and range for continuous variables and percentage for qualitative variables.

Scores were described at each measurement time with mean, standard deviation (SD), median and range, for all patients and according to the idarubicin dose level (data not shown). The difference between follow-up scores and baseline scores were also calculated and described with mean, SD and 95% confidence interval (CI), for all patients and according to the idarubicin dose level.

Longitudinal analysis

The objective was to explore added value of HRQoL to identify RP2D in a phase I clinical trial with the time to deterioration approach. Thus, we compared to time to deterioration in a HRQoL score according to idarubicin level dose. Several definitions of deterioration have been proposed integrating or not the occurrence of a severe toxicity of grade 3/4 according to the NCI-CTCAE criteria made by the clinician.

Four definitions of TTD were investigated:

5. The time from inclusion in the study to a first deterioration with a 5-point MCID as compared to the baseline score (13);

6. The time from inclusion in the study to a first deterioration with a 5-point MCID as compared to the baseline score, integrating all-cause death as event. This definition was redefined as the HRQoL-deterioration free survival (QFS)(11);

7. The TTD in a least one HRQoL score among the four targeted HRQoL dimensions. In this definition, the event corresponds to the first deterioration observed, whatever the HRQoL score deteriorated (12);

8. The TTD in at least one HRQoL score or occurrence of at least one severe toxicity of grade 3/4 according to the NCI-CTCAE criteria, whatever which event occurs first. This last definition allows to integrate the toxicity according to the NCI-CTCAE scale in the TTD definition.

Patients with no baseline score were censored at baseline. Patients with a baseline score but with no follow-up score were censored just after baseline (one day after baseline).

Sensitivity analyses were conducted for each definition of TTD considering patients with no baseline HRQoL score and those with no follow-up HRQoL score as event.

TTD curves were calculated using the Kaplan-Meier estimation method and described using median and 95% CI. TTD were compared according to the idarubicin dose level using for exploratory purposes using log-rank tests and described using Univariate Hazard Ratio (HR) with 95% CI: Univariate HR of 10 mg idarubicin vs. 5 mg idarubicin dose level (10 vs. 5 mg) and Univariate HR of 15 mg idarubicin vs. 5 mg idarubicin dose level (15 vs. 5 mg) were estimated.
Univariate Cox regression analyses were conducted as exploratory analyses in order to investigate factors potentially influencing the TTD for the two last definitions:

- the TTD in at least one HRQoL score;

- the TTD in at least one HRQoL score or occurrence of at least one grade 3/4 toxicity.

Variables tested were the gender (women vs. men), age (continuous variable), the ECOG PS (1 vs. 0), the Child-Pugh class (A6 vs. A5 and B7 vs.A5), CLIP score (2

vs. 1 vs. 0), the BCLC stage (B vs. A; C vs. A) and the occurrence of a DLT (yes vs. no). Regarding the definition of the TTD in at least one HRQoL score, the time to grade 3/4 toxicity was also tested (time dependent variable) in order to test the impact of the occurrence of the grade 3/4 toxicity on HRQoL.

All analyses were performed with R software (22).

RESULTS

Study population

Between March 2010 and March 2012, 21 patients were included: 9, 6, and 6 patients were treated at idarubicin dose levels of 5-, 10-, and 15-mg, respectively (Fig 1). The median age was 64 years [range 45-79] and 18 patients were men (86%). 20 patients have a good ECOG PS (95%). Baseline characteristics of patients are summarized in Table 1.

A DLT was observed for three patients: the patient n°1 included at 10-mg idarubicin dose level experienced a DLT 27 days after TACE (hyperbilirubinemia); the patients n°3 included at 5-mg experienced a DLT 24 days after TACE (grade 4 cardiac event); the patient n°21 included at 15-mg idarubicin level experienced a DLT two days after TACE (transient elevated aspartate aminotransferase). Calculated MTD of idarubicin was 10 mg.

Descriptive analysis

Table 2 contains the mean difference between follow-up scores and baseline scores according to idarubicin level dose.

Between follow-up 15 days after TACE and baseline:

- GHS score decreases with a mean difference of 5.6 [CI95% -17.5; 6.4], 16.7 [-58.1; 27.7] and 8.3 [-21.0; 4.3] points for 5-, 10- and 15-mg idarubicin dose level respectively

- Fatigue score increases with a mean difference of 18.5 [-45.2; 82.3], 29.6 [-34.1; 93.4] and 37.8 [6.1; 69.5] points for 5-, 10- and 15-mg idarubicin dose level respectively.

One month after TACE, fatigue score increases with a mean difference of 14.8 [-42.6; 72.3], 18.9 [-11.8; 49.6] and 9.3 [-7.9; 26.4] points for 5-, 10- and 15-mg idarubicin dose level respectively.

Two months after TACE, fatigue score increases for 5-mg and 10-mg idarubicin dose level with a mean difference of 22.2 [22.2-22.2] and 2.6 [-12.1; 23.2] points respectively and decreases of 5.6 points [-36.2; 25.1] for 15-mg idarubicin dose level.

Longitudinal analysis

Table 3 contains the results of the TTD and QFS analyses according to idarubicin level dose for the four targeted dimensions of HRQoL with the sensitivity analysis.

TTD with a 5-point MCID of GHS

Eight patients presented a significant deterioration of the GHS with a 5-point MCID among the 21 included patients: 2 patients at 5 mg idarubicin dose level, 2 patients at 10 mg and 4 patients at 15 mg with a median TTD of 19 days [18-NA], 23 days [20-NA] and 25 days [17-NA] respectively. The Univariate HR 10 vs. 5 mg is 0.50 [0.07-3.59]. The Univariate HR 15 vs. 5 mg is 0.85 [0.15-4.71].

Integrating all-cause death as event, 10 patients presented a significant deterioration of the GHS with a 5-point MCID or death among the 21 included patients: 3 patients at 5 mg idarubicin dose level, 3 patients at 10 mg and 4 patients at 15 mg with a median QFS of 19 days [18-NA], 140 days [20-NA] and 25 days [17-NA] respectively. The Univariate HR 10 vs. 5 mg is 0.87 [0.14-5.36]. The Univariate HR 15 vs. 5 mg is 1.17 [0.19-7.14].

TTD with a 5-point MCID of physical functioning

Six patients presented a significant deterioration of the physical functioning with a 5-point MCID among the 21 included patients: 1 patient at 5 mg idarubicin dose level, 1 patient at 10 mg and 4 patients at 15 mg with a median TTD of NA days [19-NA], NA days [23-NA] and 61 days [14-NA] respectively. The Univariate HR 10 vs. 5 mg is 0.68 [0.04-10.84]. The Univariate HR 15 vs. 5 mg is 1.70 [0.18-16.38].

Integrating all-cause death as event, 10 patients presented a significant deterioration of the physical functioning with a 5-point MCID or death among the 21 patients included: 3 patients at 5 mg idarubicin dose level, 3 patients at 10 mg and 4 patients at 15 mg with a median QFS of 529 days [19-NA], 256 days [23-NA] and 61 days [14-NA] respectively. The Univariate HR 10 vs. 5 mg is 0.67 [0.11-4.13]. The Univariate HR 15 vs. 5 mg is 2.76 [0.40-19.27].

TTD with a 5-point MCID of fatigue

Ten patients presented a significant deterioration of the fatigue with a 5-point MCID among the 21 included patients: 3 patients at 5 mg idarubicin dose level, 3 patients at 10 mg and 4 patients at 15 mg with a median TTD of 19 days [18-NA], 29 days [20-NA] and 19 days [14-NA] respectively. The Univariate HR 10 vs. 5 mg is 0.77 [0.13-4.71]. The Univariate HR 15 vs. 5 mg is 1.21 [0.22-6.69].

The same results are obtained for QFS of fatigue.

TTD with a 5-point MCID of pain

Seven patients presented a significant pain deterioration with a 5-point MCID among the 21 included patients: 2 patients at 5 mg idarubicin dose level, 1 patient at 10 mg and 4 patients at 15 mg with a median TTD of 76 days [39-NA], NA days [23-NA] and 26 days [14-NA] respectively. The Univariate HR 10 vs. 5 mg is 0.77 [0.05-12.36]. The Univariate HR 15 vs. 5 mg is 3.89 [0.42-36.06].

Integrating all-cause death as event, 10 patients presented a significant pain deterioration with a 5-point MCID or death among the 21 patients included: 3 patients

at 5 mg idarubicin dose level, 3 patients at 10 mg and 4 patients at 15 mg with a median QFS of 76 days [39-NA], 256 days [23-NA] and 26 days [14-NA] respectively. The Univariate HR 10 vs. 5 mg is 0.52 [0.09-3.16]. The Univariate HR 15 vs. 5 mg is 3.26 [0.49-21.54].

TTD in at least one HRQoL score integrating or not toxicity as an event

Twelve patients presented a significant deterioration of at least one HRQoL score with a 5-point MCID among the 21 included patients: 3 patients at 5 mg idarubicin dose level, 3 patients at 10 mg and 6 patients at 15 mg with a median TTD of 19 days [18-NA], 29 days [20-NA] and 17 days [14-NA] respectively. The Univariate HR 10 vs. 5 mg is 0.78 [0.13-4.68]. The Univariate HR 15 vs. 5 mg is 2.99 [0.56-15.91].

Thirteen patients presented a significant deterioration of at least one HRQoL score with a 5-point MCID or a toxicity of grade ¾ among the 21 included patients: 3 patients at 5 mg idarubicin dose level, 4 patients at 10 mg and 6 patients at 15 mg with a median TTD of 6 days [5-NA], 8 days [5-NA] and 6 days [1-NA] respectively. The Univariate HR 10 vs. 5 mg is 0.91 [0.2-4.12]. The Univariate HR 15 vs. 5 mg is 1.44 [0.34-6.13].

Whatever the definition of deterioration applied, patients included at 10-mg and 15-mg idarubicin dose level presented a longer TTD than those included at 5-mg for sensitivity analyses considering patients with no baseline score and those with no follow-up score as events.

Univariate Cox Analyses for TTD of at least one HRQoL score integrating or not toxicity as an event

Women presented a shorter TTD in at least one HRQoL score than men (HR 14.7 [1.31-164.7]). Considering no follow-up as event, older patients presented a longer TTD than younger one (HR 0.93 [0.88 - 0.99]) with a *P*-value=0.03.

Regarding the TTD in at least one HRQoL score or a toxicity of grade 3/4, women and younger patients were both independently associated with shorter TTD with Univariate HR equals to 11.01 [1.47 - 82.48] for women vs. men and 0.91 [0.83 - 0.99] for age respectively.

DISCUSSION

The study investigated added value of HRQoL using TTD approach according to the idarubicin dose level in a phase I clinical trial of HCC. HRQoL results are coherent and in the same way to the 10-mg idarubicin dose level retained as the MTD. Patients at the 10-mg idarubicin dose level presented a trend of longer time to GHS, physical functioning, fatigue and pain deterioration whatever the definition of deterioration applied. Regarding the TTD with a 5-point MCID, the Univariate HR of 10-mg idarubicin dose level as compared 5-mg was 0.50 [CI 95% 0.7-3.59], 0.68 [0.0-10.84], 0.77 [0.13-4.71] and 0.77 [0.05-12.36] for the GHS, physical functioning, fatigue and pain respectively. The composite definition of deterioration integrating toxicity according to the NCI-CTCAE criteria as event has the advantage to take into account the measurement of treatment tolerability made by the clinician while competitive risk could occurred. Using this definition, patients included at the 10-mg idarubicin dose level still presented a trend of longer TTD than patients included at 5-mg idarubicin dose level.

Patients included at 15-mg idarubicin dose level also presented a longer time to GHS deterioration with a 5-point MCID as compared to those included at 5-mg idarubicin dose level with a HR of 0.85 [0.15-4.71] of 15-mg vs. 5-mg dose level. The sensitivity analysis integrating patients with no baseline and those with no follow-up as events highlights a longer time to GHS, physical functioning, fatigue and pain of patients at 15-mg idarubicin dose level as compared to those included at 5-mg level dose. Regarding the TTD in at least one HRQoL score integrating or not toxicity as an event, patients included at 15-mg idarubicin dose level presented a shorter TTD than those included at 5-mg but a longer TTD for the corresponding sensitivity analysis. In this way, for many definitions of TTD, patients included at 15-mg idarubicin dose level

presented a longer TTD than those included at 5-mg. Thus these findings suggest that if HRQoL would have been taken into account in the definition of DLT, the MTD might be the 15-mg idarubicin dose level.

The NCI CTC AE scale was initially developed to cytotoxic chemotherapies administered in a limited number of chemotherapy cycles including rest period. MTA are generally given during some long periods and with continuous schedules. Some moderate toxic side effects persisting in a long-period have been shown to be frequent and are not considered as a DLT (1) even if they affect patients' daily life. As example, moderate toxicities like diarrhea of grade 2 is generally not considered by the NCI-CTCAE and integrated in DLT definition (generally grade 3 at least to be considered as a DLT) while it can affect patients' daily life if this symptom persists in a long-period. Moreover, the assessment made by the clinicians generally results in an under or over-estimation of patients' side effects (23) and the gold standard could be the patients assessment instead of clinicians (24). In this context, the HRQoL assessment as well as other patient-reported outcomes could be very helpful in a phase I clinical trial in order to take into account patients' perception about the tolerability of the treatment received (2).

In this study, HRQoL was study until two months after TACE while DLT was defined according to toxicities observed until one month after TACE. It seems important to take into account treatment side effects and impact on patients' HRQoL at long term.

To our knowledge, this study is the first to investigate the TTD in a HRQoL score in the framework of oncology phase I clinical trial. Several studies investigated HRQoL in phase I clinical trials but results are more often descriptive and poorly exploited regarding the primary objective of phase I trial (9). To our knowledge, none study compared the HRQoL score according to molecular dose level allocated to the patients, or according to the occurrence or not of a toxicity or a DLT.

In this study, the QLQ-C30 questionnaire was used to assess patients' HRQoL level. However, although it is a validated tool to assess patients' HRQoL level in oncology, the QLQ-C30 may not be well adapted to patients included in a phase I clinical trial. Cox et al. investigated both interviews and questionnaires to assess patients' HRQoL in the context of cancer phase I and II clinical trials participation with the use of the QLQ-C30 questionnaire (6).

The EORTC QLQ-C15PAL questionnaire developed for patients in palliative care could be more adapted to patients in a phase I clinical trial since it was dedicated to cancer patient with similar clinical characteristics of patients (25). However, this questionnaire is a reduce form of the QLQ-C30 and contains 15 items of the QLQ-C30 questionnaire. The development of a specific patent-reported outcomes questionnaire for phase I oncology clinical trials could thus be necessary in order to capture all relevant information regarding these specific category of patients and more focused on symptoms.

This study suggests the added value of patients-reported symptoms and HRQoL of patients to complement the usual toxicity assessment as well as the DLT definition. Moreover, an increase in patients' HRQoL could enhance a better compliance with treatment (26).

To conclude, in this study we have shown that patients included at 10-mg idarubicin dose level corresponding to the MTD presented a longer TTD than other patients whatever the definition of deterioration applied. Moreover, this study highlights the necessity to complement the definition of the DLT incorporating patient-reported outcomes.

REFERENCES

1. Paoletti X, Le Tourneau C, Verweij J, Siu LL, Seymour L, Postel-Vinay S, et al. Defining dose-limiting toxicity for phase 1 trials of molecularly targeted agents: results of a DLT-TARGETT international survey. Eur J Cancer. 2014;50:2050-6.
2. Postel-Vinay S, Arkenau HT, Olmos D, Ang J, Barriuso J, Ashley S, et al. Clinical benefit in Phase-I trials of novel molecularly targeted agents: does dose matter? British journal of cancer. 2009;100:1373-8.
3. Le Tourneau C, Razak AR, Gan HK, Pop S, Dieras V, Tresca P, et al. Heterogeneity in the definition of dose-limiting toxicity in phase I cancer clinical trials of molecularly targeted agents: a review of the literature. European journal of cancer. 2011;47:1468-75.
4. Basch E. Patient-reported outcomes in drug safety evaluation. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2009;20:1905-6.
5. Verweij J, Disis ML, Cannistra SA. Phase I studies of drug combinations. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2010;28:4545-6.
6. Cox K. Assessing the quality of life of patients in phase I and II anti-cancer drug trials: interviews versus questionnaires. Social science & medicine. 2003;56:921-34.
7. Beitz J, Gnecco C, Justice R. Quality-of-life end points in cancer clinical trials: the U.S. Food and Drug Administration perspective. Journal of the National Cancer Institute Monographs. 1996:7-9.
8. Bruner DW. Should patient-reported outcomes be mandatory for toxicity reporting in cancer clinical trials? Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2007;25:5345-7.
9. Stephenson CM, Levin RD, Spector T, Lis CG. Phase I clinical trial to evaluate the safety, tolerability, and pharmacokinetics of high-dose intravenous ascorbic acid in patients with advanced cancer. Cancer chemotherapy and pharmacology. 2013;72:139-46.
10. Rouanne M, Massard C, Hollebecque A, Rousseau V, Varga A, Gazzah A, et al. Evaluation of sexuality, health-related quality-of-life and depression in advanced

cancer patients: a prospective study in a Phase I clinical trial unit of predominantly targeted anticancer drugs. European journal of cancer. 2013;49:431-8.
11. Anota A, Hamidou Z, Paget-Bailly S, Chibaudel B, Bascoul-Mollevi C, Auquier P, et al. Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization? Quality of Life Research. 2013.
12. Bonnetain F, Dahan L, Maillard E, Ychou M, Mitry E, Hammel P, et al. Time until definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. European journal of cancer. 2010;46:2753-62.
13. Hamidou Z, Dabakuyo TS, Mercier M, Fraisse J, Causeret S, Tixier H, et al. Time to deterioration in quality of life score as a modality of longitudinal analysis in patients with breast cancer. The oncologist. 2011;16:1458-68.
14. Boulin M, Hillon P, Cercueil JP, Bonnetain F, Dabakuyo S, Minello A, et al. Idarubicin-loaded beads for chemoembolisation of hepatocellular carcinoma: results of the IDASPHERE phase I trial. Alimentary pharmacology & therapeutics. 2014.
15. NCI. Common Terminology Criteria for Adverse Events v3.0. 2006.
16. Bruix J, Sherman M, Llovet JM, Beaugrand M, Lencioni R, Burroughs AK, et al. Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference. European Association for the Study of the Liver. Journal of hepatology. 2001;35:421-30.
17. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. Biometrics. 1990;46:33-48.
18. O'Quigley J, Shen LZ. Continual reassessment method: a likelihood approach. Biometrics. 1996;52:673-84.
19. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute. 1993;85:365-76.
20. Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, AobotEQoLG. B. EORTC QLQ-C30 Scoring Manual (3rd edition). Brussels: EORTC 2001 ed2001. 2001.

21. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 1998;16:139-44.
22. Team RDC. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. 2010 [cited; Available from: URL http://www.R-project.org/
23. Atkinson TM, Li Y, Coffey CW, Sit L, Shaw M, Lavene D, et al. Reliability of adverse symptom event reporting by clinicians. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2012;21:1159-64.
24. Basch E. The missing voice of patients in drug-safety reporting. The New England journal of medicine. 2010;362:865-9.
25. Groenvold M, Petersen MA, Aaronson NK, Arraras JI, Blazeby JM, Bottomley A, et al. The development of the EORTC QLQ-C15-PAL: a shortened questionnaire for cancer patients in palliative care. European journal of cancer. 2006;42:55-64.
26. McCarberg BH, Barkin RL. Long-acting opioids for chronic pain: pharmacotherapeutic opportunities to enhance compliance, quality of life, and analgesia. American journal of therapeutics. 2001;8:181-6.

Fig 1: Consort diagram

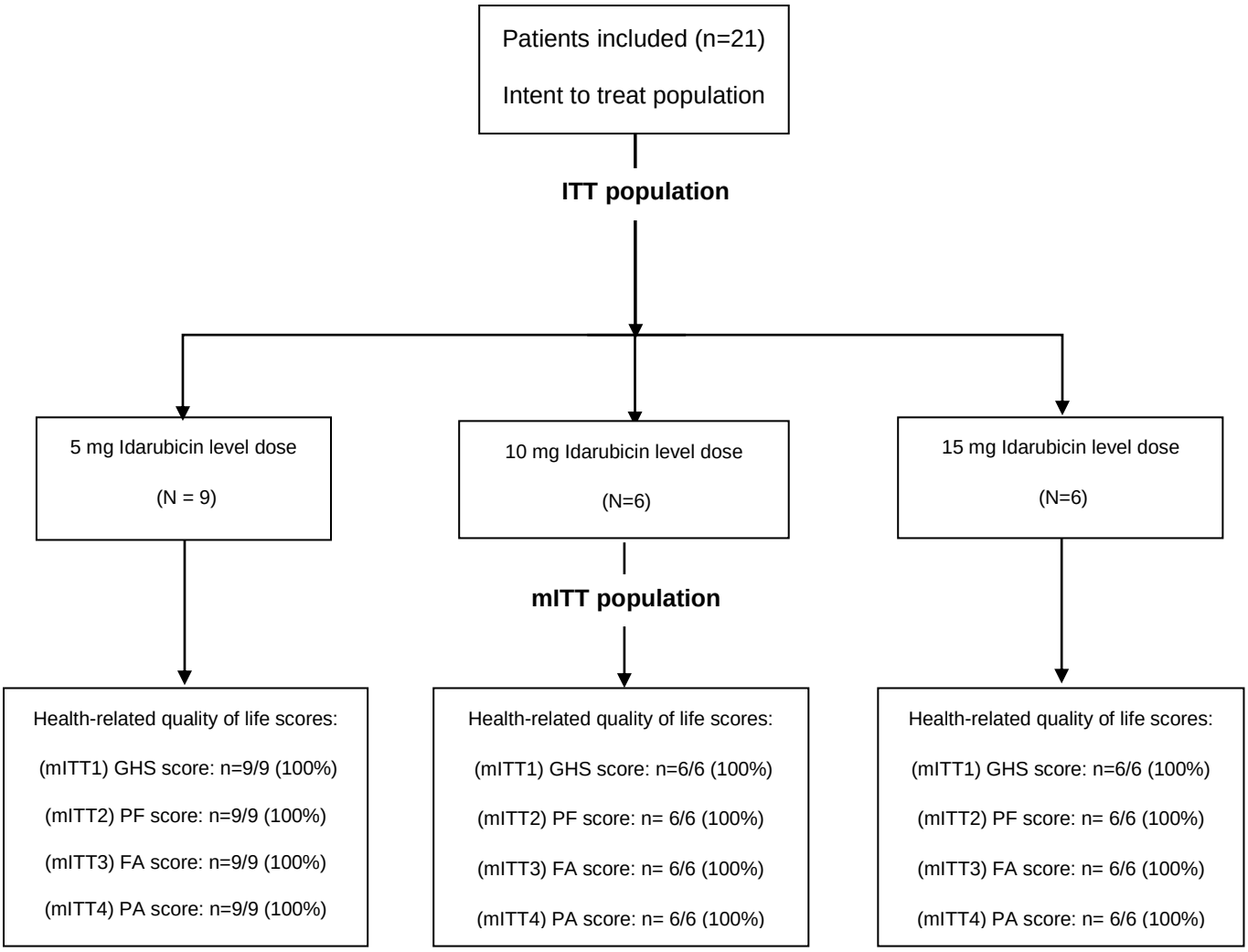


Table 1: Baseline Characteristics of patients (N=21)

Characteristic	median (range)	N (%)
Age	64 (45-79)	
Sex		
Female		3 (14)
Male		18 (86)
Etiology		
Alcohol		12 (57)
Hepatitis C infection		4 (19)
Other		5 (24)
Child-Pugh class		
A5		10 (48)
A6		6 (28)
B7		5 (24)
BCLC stage		
A		5 (24)
B		15 (71)
C		1 (5)
ECOG performance status		
0		20 (95)
1		1 (5)
Number of nodules	2 (1-6)	
Diameter of largest nodule, cm	46 (15-96)	

Table 2: Mean difference between follow-up quality of life measures and baseline measure with 95% confidence interval according to idarubicin level dose

	T1 - T0			T2-T0			T3 - T0	
	N	mean (SD)	CI 95%	N	mean (SD)	CI 95%	N	mean (SD)	CI 95%
GHS									
5 mg	3	-5.6 (4.8)	[-17.5; 6.4]	3	0.0 (14.4)	[-35.9; 35.9]	3	2.8 (4.8)	[-9.2; 14.7]
10 mg	3	-16.7 (16.7)	[-58.1; 27.7]	5	3.3 (13.9)	[-14.0; 20.6]	4	16.7 (13.6)	[-5.0; 38.3]
15 mg	5	-8.3 (10.2)	[-21.0; 4.3]	6	6.9 (13.4)	[-7.1; 21.0]	4	14.6 (12.5)	[-5.3; 34.5]
PF									
5 mg	3	-2.2 (3.8)	[-11.8; 7.3]	3	2.2 (7.7)	[-17.0; 21.3]	3	8.9 (13.9)	[-25.6; 43.4]
10 mg	3	8.9 (19.2)	[-38.9; 56.7]	5	2.7 (21.4)	[-23.4; 29.2]	4	18.3 (12.6)	[-1.7; 38.4]
15 mg	5	-5.3 (8.7)	[-16.1; 5.5]	6	-4.4 (5.4)	[-10.2; 1.3]	4	0.0 (5.4)	[-8.7; 8.7]
FA									
5 mg	3	18.5 (25.7)	[-45.2; 82.3]	3	14.8 (23.1)	[-42.6; 72.3]	3	22.2 (0.0)	[22.2; 22.2]
10 mg	3	29.6 (25.7)	[-34.1; 93.4]	5	18.9 (24.7)	[-11.8; 49.6]	4	2.6 (11.1)	[-12.1; 23.2]
15 mg	5	37.8 (25.6)	[6.1; 69.5]	6	9.3 (16.4)	[-7.9; 26.4]	4	-5.6 (19.2)	[-36.2; 25.1]
PA									
5 mg	3	0.0 (0.0)	[0.0; 0.0]	3	0.0 (16.7)	[-41.4; 41.4]	3	0.0 (16.7)	[-41.4; 41.4]
10 mg	3	-11.1 (38.5)	[-106.7; 84.5]	5	-6.7 (25.3)	[-38.1; 24.7]	4	-20.8 (25.0)	[-60.6; 18.9]
15 mg	5	23.3 (27.9)	[-11.3; 58.0]	6	2.8 (6.8)	[-4.4; 9.9]	4	4.2 (8.3)	[-9.1; 17.4]

GHS: global health status; PF: physical functioning; FA: fatigue; PA: pain

Table 3: Time to quality of life deterioration of 5 points at least as compared to the baseline score, integrating all-cause death as event (QFS) or not (TTD), with sensitivity analyses considering patients with no baseline score or with no follow up score as events

		TTD ≥ 5 point MCID			TTD ≥ 5 point MCID or no follow-up			QFS ≥ 5 point MCID			QFS ≥ 5 point MCID or no follow-up		
		N (events)	Median (CI 95%)	HR (CI 95%)	N (events)	Median (CI 95%)	HR (CI 95%)	N (events)	Median (CI 95%)	HR (CI 95%)	N (events)	Median (CI 95%)	HR (CI 95%)
GHS	All	21 (8)	23 (19-NA)		21 (16)	18 (0-NA)		21 (10)	23 (19-NA)		21 (18)	18 (0-NA)	
	5 mg	9 (2)	19 (18-NA)	1	9 (8)	0 (0-NA)	1	9 (3)	19 (18-NA)	1	9 (9)	0 (0-NA)	1
	10 mg	6 (2)	23 (20-NA)	0.50 (0.07-3.59)	6 (4)	22 (1-NA)	0.35 (0.1-1.21)	6 (3)	140 (20-NA)	0.87 (0.14-5.36)	6 (5)	22 (1-NA)	0.46 (0.15-1.44)
	15 mg	6 (4)	25 (17-NA)	0.85 (0.15-4.71)	6 (4)	25 (17-NA)	0.31 (0.09-1.06)	6 (4)	25 (17-NA)	1.17 (0.19-7.14)	6 (4)	25 (17-NA)	0.34 (0.1-1.18)
PF	All	21 (6)	92 (23-NA)		21 (14)	19 (0-NA)		21 (10)	92 (23-NA)		21 (18)	19 (0-570)	
	5 mg	9 (1)	NA (19-NA)	1	9 (7)	0 (0-NA)	1	9 (3)	529 (19-NA)	1	9 (9)	0 (0-NA)	1
	10 mg	6 (1)	NA (23-NA)	0.68 (0.04-10.84)	6 (3)	23 (1-NA)	0.37 (0.1-1.47)	6 (3)	256 (23-NA)	0.67 (0.11-4.13)	6 (5)	140 (1-NA)	0.42 (0.13-1.39)
	15 mg	6 (4)	61 (14-NA)	1.7 (0.18-16.38)	6 (4)	61 (14-NA)	0.33 (0.08-1.28)	6 (4)	61 (14-NA)	2.76 (0.4-19.27)	6 (4)	61 (14-NA)	0.48 (0.14-1.66)
FA	All	21 (10)	21 (18-NA)		21 (18)	17 (0-35)		21 (10)	21 (18-NA)		21 (18)	17 (0-35)	
	5 mg	9 (3)	19 (18-NA)	1	9 (9)	0 (0-NA)	1	9 (3)	19 (18-NA)	1	9 (9)	0 (0-NA)	1
	10 mg	6 (3)	29 (20-NA)	0.77 (0.13-4.71)	6 (5)	22 (1-NA)	0.44 (0.14-1.38)	6 (3)	29 (20-NA)	0.77 (0.13-4.71)	6 (5)	22 (1-NA)	0.44 (0.14-1.38)
	15 mg	6 (4)	19 (14-NA)	1.21 (0.22-6.69)	6 (4)	19 (14-NA)	0.37 (0.11-1.25)	6 (4)	19(14-NA)	1.21 (0.22-6.69)	6 (4)	19 (14-NA)	0.37 (0.11-1.25)
PA	All	21 (7)	76 (23-NA)		21 (15)	16 (0-NA)		21 (10)	76 (23-NA)		21 (18)	16 (0-529)	
	5 mg	9 (2)	76 (39-NA)	1	9 (8)	0 (0-NA)	1	9 (3)	76 (39-NA)	1	9 (9)	0 (0-NA)	1
	10 mg	6 (1)	NA (23-NA)	0.77 (0.05-12.36)	6 (3)	23 (1-NA)	0.39 (0.1-1.52)	6 (3)	256 (23-NA)	0.52 (0.09-3.16)	6 (5)	140 (1-NA)	0.39 (0.12-1.28)
	15 mg	6 (4)	26 (14-NA)	3.89 (0.42-36.06)	6 (4)	26 (14-NA)	0.52 (0.15-1.82)	6 (4)	26 (14-NA)	3.26 (0.49-21.54)	6 (4)	26 (14-NA)	0.52 (0.15-1.79)

*GHS: Global Health Status; PF: physical functioning; FA: fatigue; PA: pain

Table 4: Time to deterioration in at least one HRQoL score considering or not toxicity grade ¾ as event with sensitivity analysis considering patients with no follow-up as event

	deterioration in at least one HRQoL score			deterioration in at least one HRQoL score or a toxicity of grade 3/4		
	N (events)	Median days (CI 95%)	HR [CI 95%]	N (events)	Median days (CI 95%)	HR (CI 95%)
TTD						
All patients	21 (12)	20 (17-NA)		21 (13)	8 (5-47)	
5 mg	9 (3)	19 (18-NA)	1	9 (3)	6 (5-NA)	1
10 mg	6 (3)	29 (20-NA)	0.78 [0.13-4.68]	6 (4)	8 (5-NA)	0.91 [0.2-4.12]
15 mg	6 (6)	17 (14-NA)	2.99 [0.56-15.91]	6 (6)	6 (1-NA)	1.44 [0.34-6.13]
TTD or no-follow-up						
All patients	21 (20)	16 (0-23)		21 (20)	1 (1-11)	
5 mg	9 (9)	0 (0-NA)	1	9 (9)	0 (0-NA)	1
10 mg	6 (5)	22 (1-NA)	0.42 [0.14-1.32]	6 (5)	7 (1-NA)	0.43 [0.14-1.30]
15 mg	6 (6)	17 (14-NA)	0.68 [0.23-2.02]	6 (6)	6 (-NA)	0.51 [0.18-1.47]

Fig. 2: Kaplan- Meier survival curves according to idarubicin level dose for the Health-related quality of life deterioration-free survival with a 5-point MCID or death

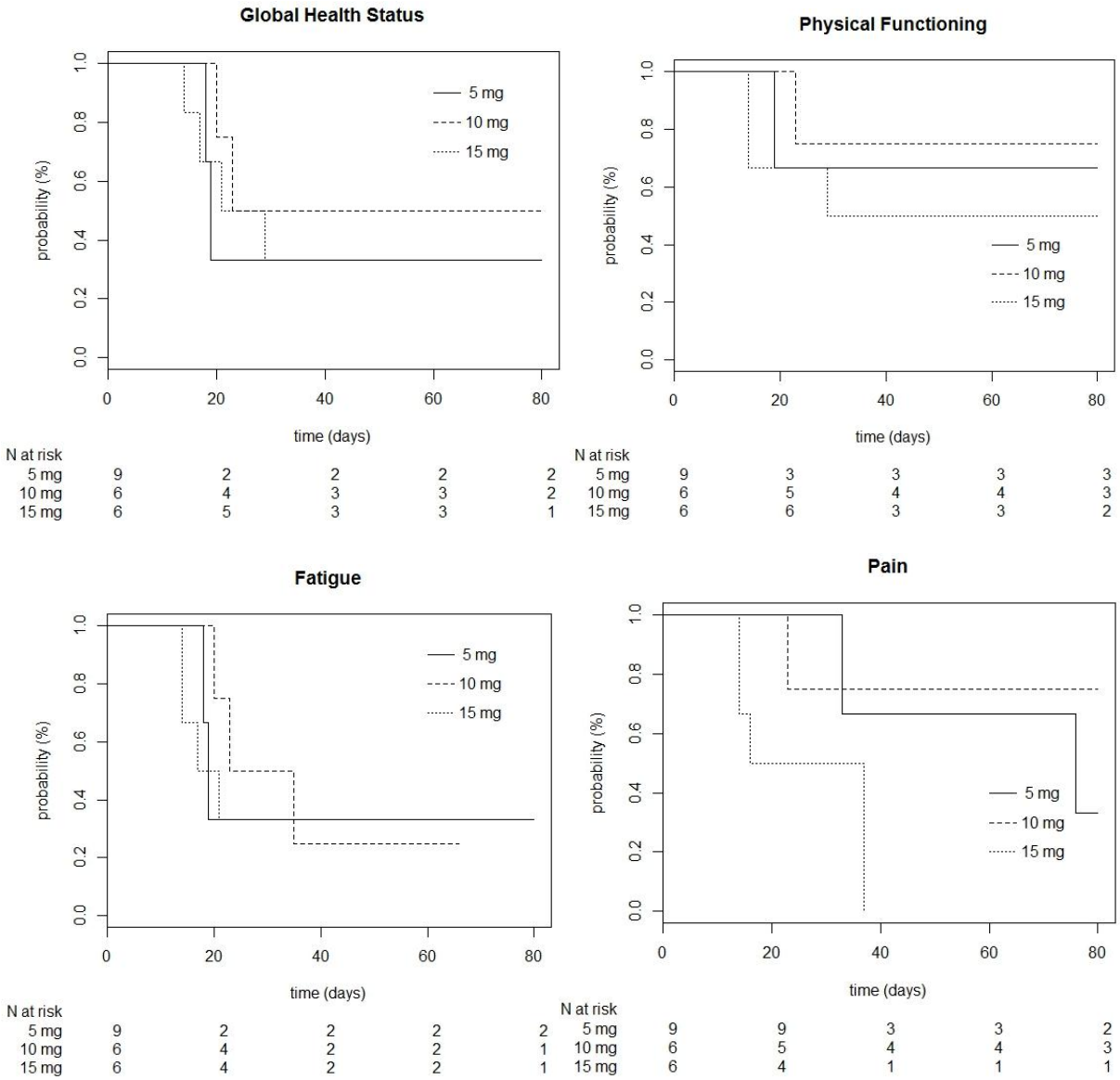


Table 5: Univariate Cox analyses for the time to deterioration in at least one score with or with not toxicity of grade ¾ as an event and sensitivity analysis integrating no follow-up as event

		MCID ≥ 5 points			MCID ≥ 5 points or no follow-up		
		N (events)	HR [CI 95%]	Log-rank test	N (events)	HR [CI 95%]	Log-rank test
TTD in at least one score							
sex	men	21 (12)	1	**0.02**	21 (20)	1	0.21
	women		14.69 [1.31 - 164.7]			2.36 [0.62 - 8.91]	
age		21 (12)	0.92 [0.84 - 1.01]	0.09	21 (20)	0.93 [0.88 - 0.99]	**0.03**
ECOG PS	0	14 (11)	1	0.52	14 (13)	1	0.73
	1		1.74 [0.35-8.68]			1.32 [0.28 - 6.27]	
Child-Pugh class	A5	15 (12)	1	0.17	15 (14)	1	0.23
	A6		2.73 [0.72-10.38]	0.14		2.58 [0.77 - 8.66]	0.12
	B7		4.46 [0.72 - 27.47]	0.11		2.97 [0.54 - 16.46]	0.21
score CLIP	0	14 (11)	1	0.25	14 (13)	1	0.20
	1		1.23 [0.27 - 5.57]	0.79		0.92 [0.23 - 3.74]	0.91
	2		4.34 [0.69 - 27.21]	0.12		3.45 [0.72 - 6.51]	0.12
BCLC stage	A	14 (11)	1	0.85	14 (13)	1	0.99
	B		0.68 [0.18 - 2.56]	0.57		0.99 [0.30 - 3.27]	0.98
	C		0.82 [0.09 - 7.22]	0.85		0.83 [0.09 - 7.30]	0.87
idarubicin level dose	5 mg	21 (12)	1	0.16	21 (20)	1	0.33
	10 mg		0.77 [0.13 - 4.68]	0.78		0.42 [0.14 - 1.32]	0.14
	15 mg		2.99 [0.56 - 15.91]	0.20		0.68 [0.23 - 2.02]	0.49
DLT occurrence	no	21 (12)	1	0.70	21 (20)	1	0.29
	yes		0.74 [0.16-3.46]			0.45 [0.10 - 1.95]	
Time to toxicity		21 (12)	3.14 [0.66 - 14.91]	0.15	21 (20)	2.48 [0.63 - 9.72]	0.19
TTD in at least one score or toxicity of grade 3/4							
sex	men	21 (13)	1	0.02	21 (20)	1	0.12
	women		11.01 [1.47 - 82.48]			2.91 [0.76 - 11.15]	
Age		21 (13)	0.91 [0.83 - 0.99]	0.04	21 (20)	0.94 [0.88 - 0.99]	0.04
ECOG PS	0	14 (12)	1	0.93	14 (13)	1	0.82
	1		0.93 [0.20 - 4.41]			0.84 [0.18 - 3.91]	
Child-Pugh class	A5	15 (13)	1	0.07	15 (14)	1	0.17
	A6		3.94 [0.93 - 16.67]	0.06		2.80 [0.74 - 10.57]	0.13
	B7		6.05 [0.90 - 40.52]	0.06		4.04 [0.68 - 24.09]	0.13
score CLIP	0	14 (12)	1	0.11	14 (13)	1	0.19
	1		1.72 [0.42 - 7.02]	0.45		1.30 [0.34 - 4.93]	0.70
	2		6.10 [1.11 - 33.37]	0.04		3.76 [0.80 - 17.75]	0.09
BCLC stage	A	14 (12)	1	0.91	14 (13)	1	0.99
	B		0.75 [0.21 - 2.64]	0.66		0.91 [0.27 - 3.02]	0.88
	C		0.90 [0.10 - 7.87]	0.92		0.88 [0.10 - 7.71]	0.91
idarubicin level dose	5 mg	21 (13)	1	0.77	21 (20)	1	0.25
	10 mg		0.91 [0.20 - 4.12]	0.91		0.43 [0.14 - 1.30]	0.13
	15 mg		1.44 [0.34 - 6.13]	0.62		0.51 [0.18 - 1.47]	0.21
DLT occurrence	no	21 (13)	1	0.92	21 (20)	1	0.55
	yes		1.07 [0.58 - 4.02]			0.68 [0.20 - 2.38]	

3. Comparaison de trois méthodes statistiques pour l'analyse longitudinale de la QdV par le biais de simulations

Résumé

Introduction

La majorité des essais cliniques intègrent désormais la qualité de vie relative à la santé (QdV) comme un critère de jugement majeur afin d'investiguer le bénéfice clinique de nouvelles stratégies thérapeutiques pour le patient et le système de santé. Cependant, l'analyse longitudinale de la QdV reste complexe et non standardisée. De plus, il parait nécessaire de proposer des méthodes statistiques accessibles et des résultats compréhensibles et pertinents pour le clinicien. L'objectif est de comparer, d'un point de vue statistique, trois stratégies d'analyses longitudinales pour les données de QdV dans les essais cliniques en cancérologie par le biais d'une étude de simulation.

Méthodes

Les méthodes proposées étaient le modèle linéaire à effets mixtes basé sur le score de QdV, le modèle à crédit partiel longitudinal (LPCM) issu de la théorie de réponse à l'item et une approche d'analyse de survie basée sur le temps jusqu'à détérioration d'un score de QdV (TJD). Ces méthodes ont été comparées par simulations selon l'erreur de type I et la puissance statistique du test d'un effet d'interaction entre le bras de traitement et le temps. Les données de QdV longitudinales ont été simulées par le biais d'un modèle LPCM en considérant que le trait latent (ici la QdV) suit une loi normale multivariée avec une matrice de variance-covariance autorégressive d'ordre 1. Différents scénarios de simulations ont été explorés basés sur la construction des questionnaires de QdV de l'EORTC et en faisant varier le nombre de patients (100, 200 ou 300 patients), d'items (1, 2 ou 4 items) et le nombre de modalités de réponses par item (4 ou 7 modalités). Cinq ou dix temps de mesure ont été considérés avec une corrélation faible (0.4), moyenne (0.7) ou élevée (0.9) entre chaque mesure. Afin de refléter les conditions réelles des essais cliniques, l'impact de l'occurrence de données manquantes informatives de type intermittentes et monotones a également été étudié.

Résultats

En ce qui concerne les données complètes, le taux d'erreur de type I était proche de la valeur attendue (5%) pour toutes les méthodes et le modèle linéaire à effet mixte était la méthode la plus puissante. La puissance du TJD était très faible pour les dimensions à un seul item puisque seulement quatre valeurs sont possibles pour le score. Quand le nombre d'items et/ou de modalités de réponse par item augmente, la puissance statistique de la méthode du TJD augmentait alors que la puissance des autres méthodes restait stable. Le modèle LPCM étaient moins puissant lorsque 10 temps de mesure étaient simulés. En présence de données manquantes informatives, la puissance des différentes méthodes diminuait, exceptée pour le TJD définitif par rapport au meilleur score antérieur et ainsi que pour le modèle LPCM. La puissance du TJD définitif par rapport au meilleur score antérieur restait stable, voire avait tendance à augmenter. Celle du modèle LPCM restait stable pour 5 temps de mesure et augmentait avec 10 temps de mesure.

Conclusion

Les résultats obtenus montrent sans ambiguïté que le modèle linéaire à effets mixtes est la méthode la plus efficace même si elle peut être critiquée de par la nature des données brutes du questionnaire. Le modèle LPCM, quant à lui, leur correspond beaucoup mieux mais est finalement compliqué à mettre en œuvre pour une efficacité moindre. Finalement, l'approche du TJD, qui commence à être largement utilisée dans l'analyse longitudinale de la QdV, pourrait être recommandée comme optimale, selon ces résultats, pour les essais cliniques de phase III avec des échelles de QdV multi-items. Cette méthode doit cependant être évitée pour les échelles évaluées par un item du fait de son manque de puissance.

<u>Article:</u> Comparison of three longitudinal analysis models for the health-related quality of life in oncology: a simulation study

Article soumis à *Health and Quality of Life Outcomes*

Amélie Anota[1,2], Antoine Barbieri[3,4], Marion Savina[5,6], Alhousseiny Pam[2], Sophie Gourgou-Bourgade[3], Franck Bonnetain[1,2], Caroline Bascoul-Mollevi[3]

[1] Quality of Life in Oncology National Platform, France

[2] Methodological and Quality of Life in Oncology Unit, EA 3181, University Hospital of Besançon, Besançon, France

[3] Biostatistic unit, Institut régional du Cancer de Montpellier (ICM) - Val d'Aurelle, Montpellier, France

[4] Institut de Mathématiques et de Modélisation de Montpellier, University of Montpellier 2, France

[5] INSERM, Clinical and Epidemiological Research Unit (CIC-EC 7) – CTD INCa, Institut Bergonié, Bordeaux, France

[6] INSERM CIC-EC7 Axe Cancer, Université de Bordeaux, France

[§]Corresponding author

Email addresses:

AA: aanota@chu-besancon.fr

AB: Antoine.Barbieri@icm.unicancer.fr

MS: Marion.Savina@isped.u-bordeaux2.fr

AP: apam@chu-besancon.fr

SGB: Sophie.Gourgou@icm.unicancer.fr

FB: franck.bonnetain@univ-fcomte.fr

CBM: Caroline.Mollevi@icm.unicancer.fr

Abstract

Background

Health-Related Quality of Life (HRQoL) is an important endpoint in oncology clinical trials aiming to investigate the clinical benefit of new therapeutic strategies for the patient. However, the longitudinal analysis of HRQoL remains complex and unstandardized. Moreover, it is necessary to propose accessible statistical methods and meaningful results for the clinicians. The objective was to compare three strategies of longitudinal analyses for HRQoL data in oncology clinical trials through a simulation study.

Methods

The methods proposed were the score and mixed model (SM), the longitudinal partial credit model (LPCM) and survival analysis approach based on the time to HRQoL score deterioration (TTD). Simulations compared the methods regarding the type I error and statistical power of the test of an interaction effect between treatment arm and time. Several scenarios of simulations were explored based on the EORTC HRQoL questionnaires and varying the number of patients (100, 200 or 300), items (1, 2 or 4) and response category per item (4 or 7). Five or 10 measurement times were considered with a low to high correlation between each measure. The impact of informative missing data on these methods was also studied to reflect most of clinical trials.

Results

With complete data, type I error rate were closed to the expected value (5%) for all methods and SM method was the most powerful method. The power of the TTD is low for uni-item dimension because only four possible values exist for the score. When the number of items increases, the power of TTD method increases while the power of LPCM and SM remains stable. For 10 measurement times, the LPCM is less efficient. With informative missing data, the results for SM and LPCM are similar whereas the power of TTD method increases.

Conclusions

To conclude, SM method is as efficient as LPCM. Moreover, the TTD approach, which began to be used extensively in HRQoL data analysis, could be optimal for phase III clinical trials with a multi-item scale.

Keywords: longitudinal analysis, statistical methods, health-related quality of life, oncology clinical trials

Background

Health-Related Quality of Life (HRQoL) is an important endpoint in oncology clinical trials aiming to investigate the clinical benefit of new therapeutic strategies for the patient and health care system [1]. However, the longitudinal analysis of HRQoL remains complex and unstandardized. At this time, no recommendations have been made to analyze longitudinal HRQoL data in oncology which is a key issue to compare results between trials. Moreover, it is necessary to propose accessible statistical methods and meaningful results for the clinicians.

HRQoL is a subjective endpoint which is not directly observable; therefore HRQoL is considered as a latent trait. Patients' HRQoL level is generally estimated through the administration of validated questionnaires given to patients at different time points for a longitudinal approach.

In oncology clinical trials, one of the most widely used questionnaire is the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ-C30) which is a validated self-administered questionnaire specific to cancer [2]. According to cancer sites and therapeutic settings, some specific modules can be added to the QLQ-C30 such as the QLQ-BR23 for breast cancer [3]. The QLQ-C30 and all additional HRQoL modules of the EORTC HRQoL group are multidimensional questionnaires which allow to evaluate several HRQoL domains (functional and symptomatic) specific to cancer. Each dimension is evaluated through one or more items on a 4-point Likert scale (1: 'Not at all'/2: 'A little'/3: 'Quite a bit'/4: 'A lot'), except for the global health status dimension of the QLQ-C30 which is evaluated through two items on a 0-7 scale.

A score is then established for each dimension reflecting the patient's level on the corresponding scale. The score is the sum of patients' responses for a given dimension and can be calculated if at least half of the items are answered. The score is normalized on a 0-100 scale so that a high score corresponds to a high level of functioning on a functional scale or to a high incidence of symptoms on a symptomatic scale. Conversely, a low score corresponds to a low level of functioning on a functional scale or to a low incidence of symptom on a symptomatic scale. If some items remained unanswered in a given dimension, the EORTC recommends to replace the missing response by the mean of answered items scores (personal mean score imputation method) if at least half of the items are answered [4].

In oncology clinical trials, these questionnaires are administered to the patients several times depending on the therapeutic setting: generally, at baseline (before randomization), during the treatment (e.g. at each chemotherapy cycle), at the end of the study and/or repeatedly during the follow-up until tumor progression. Therefore, the objective is to analyze the evolution of patients' HRQoL level over time. Given this longitudinal assessment, data are often missing, particularly in the advanced or metastatic settings [5]. Patients may not complete one or more items in a given questionnaire (intermittent missing items) [6]. Entire forms may also be missing if the patient cannot fill out the HRQoL questionnaire at a measurement time (intermittent missing form) [7]. Finally, patients may drop-out the study prematurely, generally due to a deterioration of health state or to death (monotone missing data) [8].
Three types of missing data exist according to the classification of Little and Rubin [9]. If the missing data are not dependent on the latent trait (the HRQoL level) or on a previously observed variable, they are considered as missing completely at random (MCAR). For example, a patient can forget to complete an item or a questionnaire at one measurement time. Missing data are missing at random (MAR) if they are not dependent on the latent variable but can be explained by a previously observed variable. For example, the age of some patients may explain their reluctance to answer a particular question. Finally, missing data are missing not at random (MNAR) if they are dependent on the latent trait. For example, if the patient did not complete a questionnaire due to his/her altered health status, it can reflect a deterioration of his/her HRQoL level. MCAR and MAR missing data are non-informative and thus may not induce a bias in the analysis. In contrast, the MNAR profile corresponds to informative data and can bias the results if it is not adequately taken into account in the longitudinal analysis method. In oncology clinical trials and especially in advanced cancers, missing data are most often MNAR [10].

The longitudinal analysis of HRQoL data is generally performed according to the Classical Test Theory (CTT). In the CTT, the score constructed from the items answers is considered as a good representation of the "true" HRQoL level. Therefore, the longitudinal analysis is based on this score, considering that it is a semi-quantitative measure, even if only one item allows to construct the score. The Item Response Theory (IRT) is another approach in which items play a key role [11]. The models from IRT link the item responses to the latent trait by a probabilistic model with generally a logistic link. An important class of IRT models is the Rasch-family models [12].

Some previous studies already compared CTT and IRT approaches for the longitudinal analyses of patient-reported outcomes such as HRQoL [13-15]. These studies highlighted the similar performance of both approaches in the context of complete data [13] and in the presence of monotone missing data [14]. In the presence of informative intermittent missing data, the Rasch-family models seem more efficient than CTT and particularly provide a high statistical power [15]. However, all these studies were performed on dichotomous items and restricted to three measurement times. Dichotomous items are rarely used in HRQoL questionnaires. The EORTC HRQoL questionnaires and most of the other HRQoL questionnaires are built on a Likert scale with polytomous items. Moreover, in oncology clinical trials, more than three measurement times are generally planned. Therefore, it is necessary to compare these two approaches in the context of polytomous items and with more than three measurement times. These previous studies also focused on the effects of time or treatment arm. In randomized clinical trials, HRQoL level is supposed to be equal in both treatment arms at baseline. To detect a different effect, we study if there is a significant difference between both arms concerning the HRQoL level evolution, using an interaction parameter between treatment arm and time.

In previous studies, the evaluated CTT-based approach was the score and mixed model (SM). This method is the most widely used for longitudinal analyses. However, in oncology clinical trials, a time to event approach i.e. the so-called time to HRQoL score deterioration (TTD) began to be used extensively [16-19]. In fact, this method has the advantage to produce meaningful results for the clinicians as compared to IRT models and more generally mixed models. To date, no study has yet compared this method to the SM and IRT models.

The objective of this study was to compare these three statistical methods to analyze longitudinal HRQoL data in oncology clinical trials through a simulation study:

- two CTT-based approaches, namely the SM model and the TTD approach;
- and a longitudinal IRT model for polytomous items called the Longitudinal Partial Credit Model (LPCM).

Simulations compared the methods regarding the type I error rate and statistical power of the test of interaction effect between treatment arm and time. To reflect the reality of most clinical trials, the impact of informative missing data on these methods was also studied with the implementation of both intermittent and monotone missing data depending on patients' HRQoL level (MNAR profile).

Methods

Longitudinal analysis models for health-related quality of life

Score and mixed model

In CTT, the observed score is considered to be closed to the real HRQoL level, i.e. the relationship between the observed score and the "true" score is linear.

The SM model, based on the CTT approach, involves applying a linear mixed model on the observed scores computed at each measurement time.

The score was computed according to the recommendations of the EORTC HRQoL questionnaires for a symptomatic scale or global health status scale [4]. The score Y_n for the *n-th* patient for a dimension composed of *I* items is then equals to $\left(\left(\frac{1}{I}\sum_i X_i\right) - 1\right) \times \frac{100}{r}$ with r as the difference between the highest and the lowest possible response to the items.

We considered a model with two fixed effects: an interaction effect between treatment arm and time (difference of HRQoL evolution between both treatments) and a time effect (HRQoL evolution). Moreover, we added a random effect on patient (individual deviance from average intercept) and time (individual deviance from average time effect) with an unstructured covariance matrix Σ. The model considered can be written as follows:

$$Y_n^{(t)} = c + v * arm_n * t + \gamma * t + u_{0,n} + u_{1,n} * t + \varepsilon_n^{(t)}$$

$$(u_{0,n}, u_{1,n}) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right)$$

and $\varepsilon_n^{(t)} \sim N(0, \sigma^2)$ independent;

Where:

- $Y_n^{(t)}$ is the score of the patient n at time t,
- c is a constant,
- arm_n is the treatment arm of patient n (equal to 0 or 1),
- v is a fixed interaction effect between treatment arm and time,
- γ is a fixed time effect,
- $u_{0,n}$ is a random effect on patient n,
- $u_{1,n}$ is a random time effect on patient n,
- Σ is the variance matrix of random effects $(u_{0,n}, u_{1,n})$,
- $\varepsilon_n^{(t)}$ is the residual of patient n at time t,
- σ^2 is the residual variance.

Estimation of the parameters was done using a Maximum Likelihood method which is based on the Newton-Raphson algorithm. The model was implemented using SAS software (SAS Institute Inc., Cary, USA) with proc MIXED.

Time to health-related quality of life score deterioration

The TTD approach is also based on the observed score and relies on the definition of the minimal clinically important difference (MCID) in order to be effective from a clinical point of view. Several definitions of TTD have been proposed according to the therapeutic situation and cancer site. Events can be defined according to the chosen reference score, MCID, missing scores, including all-cause death or not. Given the multiplicity of the possible definitions of TTD, a standardization of the longitudinal analysis of HRQoL data in oncology according to the TTD approach has been proposed [20]. Thus, four main definitions have been retained in the present paper according to these recommendations.

The most intuitive definition of the TTD is the time from inclusion-randomization in the study to a first deterioration of at least one MCID unit as compared to the baseline score [21]. Patients with no deterioration before their drop-out are censored at the time of the last follow-up or the last HRQoL assessment.

The observed deterioration can be definitive or not. In the palliative setting, it is more relevant to study the time until definitive HRQoL score deterioration (TUDD) reflecting the deterioration of the patient's health status, which is stable over time and representing an absorbing state. The TUDD has been defined as the time from inclusion in the study to a first deterioration of at least one MCID unit as compared to the baseline score with no further improvement of more than one MCID unit as compared to the baseline score or if the patient dropped out after deterioration, resulting in missing data [16].

In the published definitions, the reference score is the baseline score. However, other scores can be chosen as reference score such as the best previous score. Indeed, the baseline score is not necessary the reference score for the patient in case of a change in patient's internal standard illustrating one component of a response shift effect [20, 22, 23]. Therefore, both options were retained to study their impact on this approach.

Regarding the EORTC HRQoL questionnaires, a 5-point deterioration in HRQoL scores is generally considered as the MCID [24]. The MCID was thus fixed to 5 points.

Table 1 summarizes the four definitions retained.

Furthermore, a high HRQoL score corresponds to a high level of functioning for a functional scale and to a high presence of symptom for a symptomatic scale. Therefore, deterioration was considered as a decrease of the functional scale or global health status dimension and an increase of the symptomatic scale.

In the basic TTD and TUDD approaches, intermittent missing data were ignored considering that patient's HRQoL level remained unchanged since the last available HRQoL assessment.

The TTD and TUDD estimations were calculated using the Kaplan-Meier method [25].

These definitions of TTD and TUDD were implemented using SAS software.

Longitudinal mixed partial credit model

An important family of IRT models is the Rasch-family models. Despite the interesting properties of these models such as a specific objectivity, they are still little applied for the longitudinal analysis of HRQoL data. To date, few investigations are ongoing in clinical oncology [26, 27].

The Partial Credit Model (PCM) is a Rasch-family model adapted to polytomous items [28]. The PCM models the probability for one individual n to choose the response category k among the m_j possible responses for the item j (i.e. generalized linear mixed model with a multinomial logit link function) given the latent trait θ_n and the category difficulty parameters $\delta_{j,1}, \dots, \delta_{j,m_j}$ for the item j.

$$P\left(X_{n,j} = k \middle| \theta_n, \delta_{j,1}, \dots, \delta_{j,m_j}\right) = \frac{\exp(k\theta_n - \sum_{i=1}^{k} \delta_{j,i})}{\sum_{h=1}^{m_j} \exp(h\theta_n - \sum_{i=1}^{h} \delta_{j,i})}$$

As all Rasch-family models, the PCM relies on three fundamental assumptions:

1. the unidimensionality of the latent trait, i.e. all considered items must measure the same concept or the same HRQoL dimension for example,
2. the monotonicity , i.e. the probability to choose the response category k or a higher response category is increasing in function of the latent trait,
3. and the local independence of the items conditionally to the latent trait: i.e. the items answers must be independent from each other given the latent trait.

In this study, a longitudinal extension of the PCM to mixed-effect regression model was used and called the Longitudinal PCM (LPCM). Regarding the SM model, we considered a model with two fixed effects: an interaction between treatment and time effect and a time effect.

Moreover, we added a random effect on patient and time with an unstructured covariance matrix Σ. The model considered can be written as follows:

$$P\left(X_{n,j} = k \middle| {\theta_n}^{(t)},\ \delta_{j,1}, \ldots, \delta_{j,m_j}\right) = \frac{\exp(k{\theta_n}^{(t)} - \sum_{i=1}^{k} \delta_{j,i})}{\sum_{h=1}^{m_j} \exp(h{\theta_n}^{(t)} - \sum_{i=1}^{h} \delta_{j,i})}$$

$${\theta_n}^{(t)} = \nu * arm_n * t + \gamma * t + u_{0,n} + u_{1,n} * t$$

$$(u_{0,n}, u_{1,n}) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right)$$

$\varepsilon_n^{(t)} \sim N(0, \sigma^2)$

where the latent trait is decomposed linearly with the same effects, as for the SM.
This model was implemented using SAS software with proc NLMIXED.

Simulation algorithm

Complete data

The complete datasets were simulated in two steps.

The first step corresponded to the simulation of the latent trait $\theta_n = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)'$ for 5 measurement times for example and for $n = 1, \ldots, N$ patients. This simulation was performed for each treatment arm (0/1) with $N/2$ patients per arm. The latent trait followed a multivariate normal distribution $N_5(\mu^0, \Sigma)$ with mean $\mu^0 = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)'$ for control arm (0) and first-order autoregressive covariance matrix $\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$.

In the first-order autoregressive matrix, the correlation between HRQoL measures was assumed to decrease over time [13-15]. We fixed $\sigma^2 = 1$. For the experimental arm (1), the latent trait was assumed to follow a multivariate normal distribution $N_5(\mu^1, \Sigma)$ with mean $\mu^1 = \mu^0 + \Delta$ and with the same covariance matrix. Δ represented the treatment arm effect. In case of no treatment arm effect, $\Delta = 0$, otherwise $\Delta \neq 0$.

The second step of the complete datasets simulation corresponded to the determination of the items answers. The patients' responses to the items were obtained with a LPCM in order to respect the three assumptions of the Rasch-family models [29]. Category difficulty parameters were fixed to estimated standard normal-distribution quantiles and were similar for all items.

Several scenarios of simulations were explored based on the EORTC HRQoL questionnaires and with variations of patients number (100, 200 or 300), items (1, 2 or 4) and response category per item (4 or 7).

The value of the category difficulty parameters were as follows:

- $\delta_1 = -0.7$; $\delta_2 = 0$ and $\delta_3 = 0.7$ for items with 4 response categories,
- $\delta_1 = -1$; $\delta_2 = -0.6$; $\delta_3 = -0.2$; $\delta_4 = 0.2$; $\delta_5 = 0.6$ and $\delta_6 = 1$ for items with 7 response categories.

The simulations were performed with 4 or 7 response categories per item in order to reflect the construction of the EORTC HRQoL questionnaires. Simulations with 7 response categories per item were only performed with 2 items to illustrate the Global Health Status dimension of the QLQ-C30 questionnaire and only with 200 patients.

Five or 10 measurement times were considered with a low (0.4), moderate (0.7) or high (0.9) correlation between each measure. Each scenario was simulated with a time effect equal to:

- μ^0 = (-0.4 -0.2 0 0.2 0.4) for 5 measurement times, and
- μ^0= (-0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3 0.4 0.5) for 10 measurement times.

As the mean of the latent trait increased over time, we considered that the score observed corresponded to a symptomatic scale. In this way, in the TTD approach, the deterioration was observed when the score increased.

Each scenario was performed with a treatment arm effect ($\Delta \neq 0$) or not ($\Delta = 0$). Different treatment arm effects were tested and we retained the following effects:

- $\Delta_1 = 0$ for the first measurement time $t = 1$,
- $\Delta_t = 0.4, \forall\, t > 1$.

Generation of missing data

Simulations were then repeated with missing data generated from the complete datasets.

A latent variable φ, defined as the missing data propensity, was used to simulate missing data [30].

φ followed a multinormal distribution with mean (0 0 0 0 0)' and a variance covariance matrix equal to $\begin{pmatrix} 1 & {\rho^2}_{\theta\varphi}\rho & {\rho^2}_{\theta\varphi}\rho^2 & {\rho^2}_{\theta\varphi}\rho^3 & {\rho^2}_{\theta\varphi}\rho^4 \\ {\rho^2}_{\theta\varphi}\rho & 1 & {\rho^2}_{\theta\varphi}\rho & {\rho^2}_{\theta\varphi}\rho^2 & {\rho^2}_{\theta\varphi}\rho^3 \\ {\rho^2}_{\theta\varphi}\rho^2 & {\rho^2}_{\theta\varphi}\rho & 1 & {\rho^2}_{\theta\varphi}\rho & {\rho^2}_{\theta\varphi}\rho^2 \\ {\rho^2}_{\theta\varphi}\rho^3 & {\rho^2}_{\theta\varphi}\rho^2 & {\rho^2}_{\theta\varphi}\rho & 1 & {\rho^2}_{\theta\varphi}\rho \\ {\rho^2}_{\theta\varphi}\rho^4 & {\rho^2}_{\theta\varphi}\rho^3 & {\rho^2}_{\theta\varphi}\rho^2 & {\rho^2}_{\theta\varphi}\rho & 1 \end{pmatrix}$ for 5 measures.

$\rho_{\theta\varphi}$ represented the correlation between the latent trait θ (the HRQoL level) and the latent variable φ. The probability for a patient i to present a missing item at time t depended on his missing data propensity and is defined as:

$$p_{i,t} = P\left(MD_i{}^{(t)} = 1 \,\middle|\, \varphi_i{}^{(t)}, \pi_{min}^{(t)}, \pi_{max}^{(t)}\right) = \pi_{min}^{(t)} + \left(\pi_{max}^{(t)} - \pi_{min}^{(t)}\right)\frac{exp(\varphi_i{}^{(t)})}{1 + exp(\varphi_i{}^{(t)})}$$

with $MD_i{}^{(t)} = 1$ if patient i presented a missing data at time t. $\pi_{min}^{(t)}$ and $\pi_{max}^{(t)}$ were defined respectively as the minimum and maximum individual probability to present a missing data at time t. The expected proportion of missing data then equal to $\pi^{(t)} = \frac{\pi_{min}^{(t)} + \pi_{max}^{(t)}}{2}$. We fixed $\pi_{min}^{(t)} = 0.01$ and $\pi_{max}^{(t)} = 2\,\pi^{(t)} - 0.01$.

Patient i presented a missing data at time t according to a Bernoulli distribution with $p_{i,t}$ parameter.

Only simulation of a MNAR profile was performed. Indeed, only the MNAR profile is informative and can increase the risk of bias in the longitudinal analysis. Patients with a low HRQoL level or a high symptomatic level were supposed to be more likely to present missing data. As θ represented a symptomatic HRQoL dimension, $\rho_{\theta\varphi} > 0$ because a high level for the latent trait θ represented a high symptomatic level. We fixed $\rho_{\theta\varphi} = 0.7$ to simulate a moderate informative MNAR profile.

In order to reflect most clinical trials, both intermittent and monotone missing data were simulated. For datasets with 5 measurement times, intermittent missing data were simulated on the second and third times and monotone missing data on fourth and fifth times. For datasets with 10 measurement times, intermittent missing data were simulated from the second to the sixth measure and monotone missing data from the seventh to the tenth

measure. In both cases, the first measure was complete for all patients: no missing data was generated at baseline.

Two types of intermittent missing data were considered: intermittent missing forms and intermittent missing items. Regarding intermittent missing forms, simulation of missing data was performed at each measurement time: if patient i presents a missing data at time t, then all items of the dimensions are missing for that patient at time t. For intermittent missing item, a Bernoulli distribution with $p_{i,t}$ parameter was simulated for each item. For CTT-based methods (SM and TTD), a simple imputation of missing items was performed by the mean of the answered items if at least half of the items were answered by the patients according to the recommendation of the EORTC HRQoL questionnaires (personal mean score) to estimate the score.

Analyses were first conducted with both intermittent missing forms and drop-out and then with intermittent missing items and drop-out. Analyses were conducted with a proportion $\pi^{(t)}$ of missing data at each measurement time t equal to 10%, 20% or 30%.

Studied criteria to compare the statistical methods

The type I error rate was estimated under the null hypothesis H_0 of the absence of a treatment arm effect ($\Delta = 0$). It was calculated as the proportion of rejection of H_0 under the null hypothesis.

The statistical power of the test of an interaction effect between treatment arm and time was estimated under the alternative hypothesis H_1 of the presence of a treatment arm effect ($\Delta_1 = 0$; $\Delta_t = 0.4, \forall\, t > 1$). It was calculated as the proportion of rejection of H_0 under the alternative hypothesis H_1. A Wald test and a log-rank test were used respectively for mixed models and the survival analyses based on the TTD to test the rejection of the null hypothesis. Each scenario was simulated 500 times in order to have accurate estimations of the type I error rate and statistical power.

Results

Complete data

With complete data, the type I error rate was closed to the expected value (5%) for all methods (Table 2). The SM method was the most powerful method irrespective of the parameters value of each scenario (Table 3) followed by the LPCM method: for 5 measurement times, the range for statistical power were 40% to 99% for SM method and 38% to 95% for LPCM. The power of the TTD and TUDD approaches was low, especially for uni-item dimension. For example, with $N = 300$ patients, $I = 1$ item, $\rho = 0.4$ and 5 measures, the power of the SM method, LPCM and TTD as compared to the baseline score ("TTD baseline") were around 93%, 92% and 22% respectively. When the number of items increased, the power of TTD/TUDD method increased while the power of the LPCM and SM approaches remained stable. For 10 measurement times, the LPCM method was less powerful than for 5 measurement times. For example, when $N = 300$ patients, $I = 4$ items, $\rho = 0.7$ and with 5 measurement times, the power of the LPCM method was around 79% while that of the SM method was around 96%. With 10 measurement times and the same value for all other parameters, the power of the LPCM method decreased to 52% while that of the SM method was around 99%. The power of the SM method and the TTD/TUDD approaches increased for items with 7 response categories as compared to those with 4 response categories while the power of LPCM slightly decreased. When the correlation between measures increased, the power of the LPCM and SM methods globally tended to decrease while those of the TTD/TUDD approaches tended to increase (even if the power values remained low).

Incomplete data

With intermittent missing forms and drop-out, the type I error rate was closed to the expected value (5%) for all methods whatever the proportion of missing data (Table 4). The statistical power of the test of interaction between treatment arm and time (Table 5) decreased for SM method and TTD/TUDD approaches, except for TUDD as compared to the best previous score ("TUDD best"). With 30% missing data as compared to complete case data, 5 measurement times, N = 200 patients, I = 4 items, ρ = 0.7, statistical power decreased from 81% to 76% for SM method, from 55% to 40% for "TTD baseline", from 39.4% to 28.4% for "TTD best" and from 46% to 39% for "TUDD baseline".

Regarding TUDD as compare to the best previous score ("TUDD best"), statistical power generally increased. With 30% missing data as compared to complete case data, with 5 measurement times, N = 300 patients, I = 4 items, ρ = 0.7, statistical power increased from 30% to 36% for TUDD as compare to the best previous score.
Regarding LPCM method, statistical power decreased or remained stable with 5 measurement times while it generally increased for 10 measurement times. With 10 measurement times, N = 300 patients, I = 4 items, ρ = 0.9, statistical power of LPCM method increased from 53% with complete data to 77% with 30% missing data.

With intermittent missing items and drop-out, results were closed to those with intermittent missing forms and drop out. The type I error rate still remained stable and closed to the expected value (5%) for all methods whatever the proportion of missing data generated (see Table A1 in Additional File 1). The statistical power of the test of interaction between treatment arm and time (see Table A2 in Additional File) slightly decreased for SM method and TTD/TUDD approaches, except for TUDD as compared to the beast previous score ("TUDD best"), and whatever the number of measurement times, items, response category per item and correlations between HRQoL measures. This trend was generally more pronounced as for intermittent missing form and drop out. With 30% missing data, 5 measurement times, N = 200 patients, I = 4 items, ρ = 0.7, statistical power decreased from 81% to 72% for SM method, from 55% to 28% for "TTD baseline", from 39% to 20% for "TTD best" and from 46% to 28% for "TUDD baseline".
The statistical power of LPCM method increased with intermittent missing data. This trend was generally more pronounced as for intermittent missing form and drop out. With 10 measurement times, N = 300 patients, I = 4 items, ρ = 0.9, statistical power of LPCM method increased from 53% with complete data to 78% with 30% missing data.

The figure 1 shows that the statistical power for all methods with complete data, intermittent missing forms and drop out and intermittent missing items and drop-out for N =200 patients, a moderate correlation ($\rho = 0.7$) and 20% of missing data. The statistical power of LPCM methods increased for incomplete data as compared to complete data, for I = 2 or 4 items whatever the number of measurement times. For the same parameters values, the statistical power of SM method and TTD/TUDD approach remained stable or decreased and particularly with intermittent missing item and drop-out. For I = 1 item and 5 measurement times, the statistical power of all method remained stable. For 10 measurement times, statistical power

increased in presence of intermittent missing data for TTD/TUDD approach while it decreased for SM and LPCM approaches.

Discussion

To achieve recognition of HRQoL as major endpoint in oncology clinical trials to qualify for the patient the clinical benefit of a new therapeutic strategy, guidelines for longitudinal analyses are required. Three main methods could be proposed to analyze longitudinal HRQoL data: the SM Models, a time to event approach based on the TTD and the LPCM. This study was the first to compare these approaches for the longitudinal analysis of HRQoL data, with polytomous items and more than three measurement times. Moreover, this study was the first to address the interaction effect between treatment arm and time which corresponds to randomized clinical trials conditions with no group effect at baseline. Finally, both intermittent and monotone missing data depending on patients HRQoL level (MNAR profile) were studied, thereby approaching the actual conditions of clinical trials.

The results obtained on complete data showed that the type I error rate was closed to the expected value (5%) for all methods. Moreover, the SM model was at least as powerful as the LPCM to highlight an interaction effect between treatment arm and time. The statistical power of the TTD/TUDD approaches (whatever the definition of the deterioration considered) was very low for uni-item dimension even with a large sample size. This can be explained by the only four possible values existing for the score. Indeed, we suggest to avoid such approach in case of uni-item dimension. Therefore, 6 out of 15 dimensions of the QLQ-C30 are concerned. The statistical power of the different methods compared was also influenced by the correlation between HRQoL measures (ρ parameter): when the correlation increased, the statistical power of the SM and LPCM methods generally decreased while those of TTD/TUDD approaches increased irrespective of the other parameters value. The correlation between HRQoL measures was high if the patient's HRQoL level at one time could accurately predict his/her level at the following time. This could reflect closely spaced measures, i.e. some intensive HRQoL measure as for clinical trial with a rapid evolution of patient's health status. Conversely, a low correlation between HRQoL measures could correspond to distant measures reflecting cohort study design.

With intermittent missing data (missing items or missing forms) and drop-out, the type I error rate remained closed to the expected value for all statistical methods, whatever the proportion of missing data and the scenario considered. The statistical power generally decreased for SM and TTD/TUDD approaches except for TUDD as compared to the best previous score. For this definition, the statistical power generally increased or remained stable with the simulation of missing data. This is explained by the simulation of missing data depending on HRQoL level, i.e. patients with a low HRQoL level were more likely to present missing data. Indeed, an improvement of HRQoL level was more likely to be observed (no missing data) than a deterioration and this improvement would represent the new reference score for "TUDD best". Thus, a small deterioration of 5-point at least as compared to this new reference score was not considered as a deterioration as compared to the baseline score. Finally, this deterioration was more likely to be followed by monotone missing data, involving a definitive deterioration as compared to the best previous score.

The same trends were observed for all methods regarding statistical power whatever the type of missing data considered (intermittent missing items or missing forms). However, the statistical power decreased more for SM and TTD/TUDD approaches in presence of intermittent missing items than in presence of intermittent missing forms. Regarding analyses with intermittent missing items, the score could be estimated if at least 50% of the items are answered and considering that missing items are not informative of the patient's HRQoL level. This could result in an overestimation or underestimation of the patient's HRQoL level, which induce a bias in the longitudinal analysis.
As highlighted in other studies [14, 15], these results emphasized the limitations of the personal mean score imputation method, although it is the most often used to compute the scores. Indeed, it should be avoided, particularly when the proportion of missing data is high. Regarding LPCM method, the statistical power increased more in presence of intermittent missing items than in presence of intermittent missing forms. This is due to the specific objectivity property of Rasch-family models: even with few items answered, Rasch-family models can estimate with a high accuracy of the latent trait (i.e. HRQoL) [12].

Previous studies comparing score-based approaches and a Rasch-based approach highlighted the similar performance of SM and longitudinal Rasch models in case of complete data [13] and in the presence of monotone missing data [14]. The studies also showed that Rasch-family models seem more efficient than SM models in the presence of informative

intermittent missing data [15]. In our study, we also highlighted that the statistical power of the IRT models are less affected by the presence of missing data than those of SM method. However, contrary to previous published studies, the SM method was generally more powerful in our study than the IRT model for both complete and incomplete data with informative missing data and particularly with 10 measurement times. The good results of the SM model could be explained by a bias of fixed effect estimations because the SM model does not take into account several data characteristics such as the ceiling and floor effects or asymmetric data [31, 32]. These discrepancies with the literature may be partly also due to the number of measurement times considered. The IRT models seem to be less powerful when the number of measurement times is high. Moreover, in these studies, researchers chose to proceed in two steps to construct the longitudinal IRT model: estimation of the item parameters and HRQoL latent trait for each person at each time in a first step, and then modeling of the link between the latent trait and the time thanks to a linear mixed model. Our design also integrated at least five measures thereby reflecting a longitudinal design as used in clinical trials. Moreover, polytomous items were used in our research while dichotomous items were used in the previous studies. Finally, we were interested in the interaction effect between treatment arm and time while those studies analyzed the time effect [13-15] or the group effect [33]. Then it seems crucial to pursue the researches to test the ability of these models in the context of polytomous items.

In addition, both linear and non-linear mixed models and time to event analysis were compared. The time to event (i.e. "survival") approach based on the Time to HRQoL score deterioration was relevant in the event of a quicker alteration of patients' HRQoL in one treatment arm as compared to the other, and if this difference was maintained over time (risk proportionality). Therefore, the absence of an arm effect at baseline was coherent.

Our results correspond to a particular situation - nearly ideal but theoretical - considering that items were derived from an IRT model and that the corresponding symptomatic scale followed a multivariate normal distribution with an auto-regressive covariance matrix. Therefore, an additional work is in progress to compare these methods on real data collected from several clinical trials with various therapeutic settings, cancer sites and designs.

The standardization of the longitudinal analysis of HRQoL data in oncology clinical trials is essential to enable proper comparison of the results. To date, results of HRQoL studies are still not exploited enough to lead to changes in clinical practice. It is also necessary to provide the decision makers with meaningful and easy to understand results. In this context, the

TTD/TUDD approach is attractive for the clinicians because it is based on Kaplan–Meier survival curves and hazard ratios to qualify effect size as other well-known time-to-event important outcomes in oncology (e.g. overall survival or progression-free survival). However, this approach should be used with caution and in light of these results. Moreover, as already demonstrated for progression-free survival [34], the time interval between assessments of HRQoL could influence Kaplan Meier estimation resulting in an overestimation of TUDD. Since the occurrence of the true time of HRQoL deterioration could be unknown, dedicate statistic approaches dealing with interval assessment may be proposed. It seems also essential to properly study the profile of missing data ahead and to propose a suitable method of scores imputation in case of intermittent missing items of MNAR profile. Some methods have to be develop to use in conjunction to the TTD, such as pattern mixture models for SM model [35], in order to take into account missing data of MNAR profile.

All-cause death is usually taken into account as an event and particularly in an advanced setting [16]. However, death was not integrated in our simulation algorithm, which may explain in part the low statistical power of the TTD approach. Moreover, one advantage of this method compared to the mixed models is its adaptability to different therapeutic settings (adjuvant or advanced settings) with consideration of a transient or a definitive deterioration and with or without integration of death as an event.

In conclusion, the SM model was unambiguously the most effective method although it may be criticized by the nature of the raw data of the questionnaires. The LPCM was more adapted to this type of data but was ultimately difficult to implement and with a lower efficiency. Finally, the TTD/TUDD approach, which begins to be extensively used in the longitudinal analysis of HRQoL in oncology, could be recommended as it seemed optimal for phase III clinical trials with multi-items scales.

List of abbreviations

CTT: Classical Test Theory

EORTC: European Organization for Research and Treatment of Cancer

HRQoL: Health-related quality of life

IRT: Item Response Theory

LPCM: Longitudinal Partial Credit Model

MAR: missing at random

MCAR: missing completely at random

MCID: minimal clinically important difference

MNAR: missing not at random

PCM: Partial Credit Model

SM: score and mixed model

TTD: Time to deterioration

TUDD: Time until definitive deterioration

Competing interests

The authors declare that they have no competing interests

Authors' contributions

AA designed the study, performed the statistical analyses and interpretation and written the manuscript, AB, MS, interpreted the data and drafted the manuscript, SGB design the study, FB, CBM designed the study, managed the statistical analyses, interpreted the data and review the draft. All authors read and approved the final manuscript.

Acknowledgements

This study was supported by a grant from the French Public Health Research Institute (http://IRESP.net) under the 2012 call for projects as part of the 2009-2013 Cancer Plan.

We also thank Dr Julie Courraud for her editorial assistance.

References

1. Osoba D: **Health-related quality of life and cancer clinical trials.** *Ther Adv Med Oncol* 2011, **3:**57-71.
2. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JC, et al.: **The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology.** *J Natl Cancer Inst* 1993, **85:**365-376.
3. Sprangers MA, Groenvold M, Arraras JI, Franklin J, te Velde A, Muller M, Franzini L, Williams A, de Haes HC, Hopwood P, et al: **The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: first results from a three-country field study.** *J Clin Oncol* 1996, **14:**2756-2768.
4. Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, AobotEQoLG. B: **EORTC QLQ-C30 Scoring Manual (3rd edition). Brussels: EORTC 2001 ed2001.** 2001.
5. Fairclough DL: *Design and analysis of quality of life studies in clinical trials.* CRC press; 2010.
6. Fayers PM, Curran D, Machin D: **Incomplete quality of life data in randomized trials: missing items.** *Stat Med* 1998, **17:**679-696.
7. Curran D, Molenberghs G, Fayers PM, Machin D: **Incomplete quality of life data in randomized trials: missing forms.** *Stat Med* 1998, **17:**697-709.
8. Diggle P, Kenward MG: **Informative drop-out in longitudinal data analysis.** *Applied statistics* 1994**:**49-93.
9. Little RJ, Rubin DB: **Statistical analysis with missing data.** *New York: John Wiley & Sons* 1987.
10. Troxel AB, Fairclough DL, Curran D, Hahn EA: **Statistical analysis of quality of life with missing data in cancer clinical trials.** *Stat Med* 1998, **17:**653-666.
11. De Ayala RJ: **The theory and practice of item response theory. New York : Guilford Press,.** 2009.
12. Fischer GH, Molenaar IW: *Rasch models: Foundations, recent developments, and applications.* Springer; 1995.

13. Blanchin M, Hardouin JB, Le Neel T, Kubis G, Blanchard C, Mirallie E, Sebille V: **Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes.** *Stat Med* 2011, **30:**825-838.
14. Blanchin M, Hardouin J-B, Le Neel T, Kubis G, Sebille V: **Analysis of longitudinal Patient-Reported Outcomes with informative and non-informative dropout: Comparison of CTT and Rasch-based methods.** *International Journal of Applied Mathematics & Statistics [Internet]* 2011, **24:**I-11.
15. de Bock E, Hardouin JB, Blanchin M, Le Neel T, Kubis G, Bonnaud-Antignac A, Dantan E, Sebille V: **Rasch-family models are more valuable than score-based approaches for analysing longitudinal patient-reported outcomes with missing data.** *Stat Methods Med Res* 2013.
16. Bonnetain F, Dahan L, Maillard E, Ychou M, Mitry E, Hammel P, Legoux JL, Rougier P, Bedenne L, Seitz JF: **Time until definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma.** *Eur J Cancer* 2010, **46:**2753-2762.
17. Burris HA, 3rd, Lebrun F, Rugo HS, Beck JT, Piccart M, Neven P, Baselga J, Petrakova K, Hortobagyi GN, Komorowski A, et al: **Health-related quality of life of patients with advanced breast cancer treated with everolimus plus exemestane versus placebo plus exemestane in the phase 3, randomized, controlled, BOLERO-2 trial.** *Cancer* 2013, **119:**1908-1915.
18. Gourgou-Bourgade S, Bascoul-Mollevi C, Desseigne F, Ychou M, Bouche O, Guimbaud R, Becouarn Y, Adenis A, Raoul JL, Boige V, et al: **Impact of FOLFIRINOX compared with gemcitabine on quality of life in patients with metastatic pancreatic cancer: results from the PRODIGE 4/ACCORD 11 randomized trial.** *J Clin Oncol* 2013, **31:**23-29.
19. Kabbinavar FF, Wallace JF, Holmgren E, Yi J, Cella D, Yost KJ, Hurwitz HI: **Health-related quality of life impact of bevacizumab when combined with irinotecan, 5-fluorouracil, and leucovorin or 5-fluorouracil and leucovorin for metastatic colorectal cancer.** *Oncologist* 2008, **13:**1021-1029.
20. Anota A, Hamidou Z, Paget-Bailly S, Chibaudel B, Bascoul-Mollevi C, Auquier P, Westeel V, Fiteni F, Borg C, Bonnetain F: **Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization?** *Quality of Life Research* 2013.

21. Hamidou Z, Dabakuyo TS, Mercier M, Fraisse J, Causeret S, Tixier H, Padeano MM, Loustalot C, Cuisenier J, Sauzedde JM, et al: **Time to deterioration in quality of life score as a modality of longitudinal analysis in patients with breast cancer.** *Oncologist* 2011, **16:**1458-1468.
22. Schwartz CE, Sprangers MA: **Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research.** *Soc Sci Med* 1999, **48:**1531-1548.
23. Hamidou Z, Dabakuyo-Yonli TS, Guillemin F, Conroy T, Velten M, Jolly D, Causeret S, Graesslin O, Gauthier M, Mercier M, Bonnetain F: **Impact of response shift on time to deterioration in quality of life scores in breast cancer patients.** *PLoS One* 2014, **9:**e96848.
24. Osoba D, Rodrigues G, Myles J, Zee B, Pater J: **Interpreting the significance of changes in health-related quality-of-life scores.** *J Clin Oncol* 1998, **16:**139-144.
25. Goel MK, Khanna P, Kishore J: **Understanding survival analysis: Kaplan-Meier estimate.** *Int J Ayurveda Res* 2010, **1:**274-278.
26. Douglas JA: **Item response models for longitudinal quality of life data in clinical trials.** *Stat Med* 1999, **18:**2917-2931.
27. Glas CA, Geerlings H, van de Laar MA, Taal E: **Analysis of longitudinal randomized clinical trials using item response models.** *Contemp Clin Trials* 2009, **30:**158-170.
28. Masters GN: **A Rasch model for partial credit scoring.** *Psychometrika* 1982, **47:**149-174.
29. Holland PW, Hoskens M: **Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test.** *Psychometrika* 2003, **68:**123-149.
30. Sijtsma K, Hemker BT: **A taxonomy of IRT models for ordering persons and items using simple sum scores.** *Journal of Educational and Behavioral Statistics* 2000, **25:**391-415.
31. Hedeker D: **Multilevel models for ordinal and nominal variables.** In *Handbook of multilevel analysis.* Springer; 2008: 237-274
32. Hedeker D, Gibbons RD: *Longitudinal data analysis.* John Wiley & Sons; 2006.
33. de Bock E, Hardouin JB, Blanchin M, Le Neel T, Kubis G, Sebille V: **Assessment of score- and Rasch-based methods for group comparison of longitudinal patient-**

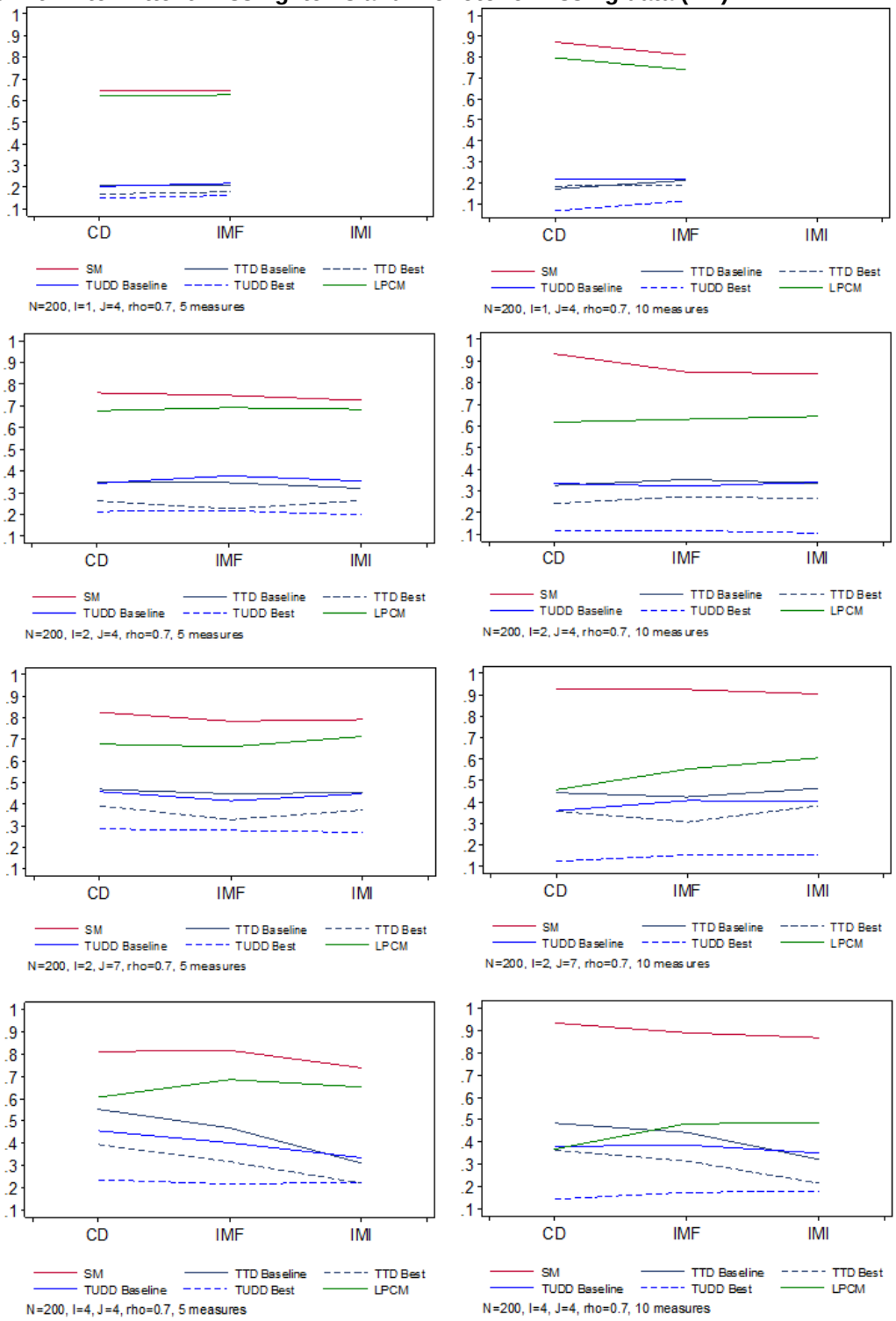
reported outcomes with intermittent missing data (informative and non-informative). *Qual Life Res* 2014.

34. Panageas KS, Ben-Porat L, Dickler MN, Chapman PB, Schrag D: **When you look matters: the effect of assessment schedule on progression-free survival.** *J Natl Cancer Inst* 2007, **99:**428-432.

35. Pauler DK, McCoy S, Moinpour C: **Pattern mixture models for longitudinal quality of life studies in advanced stage disease.** *Stat Med* 2003, **22:**795-809.

Figures Legends

Figure 1 - Power of the test of interaction between treatment arm and time for complete datasets (CD), datasets with intermittent missing forms and monotone missing data (IMF) and datasets with intermittent missing items and monotone missing data (IMI)



Methods compared are the Score and Mixed Model (SM), longitudinal Partial Credit Model (LPCM), Time to HRQoL score deterioration as compared to the baseline score (TTD baseline) or the best previous score (TTD best) and time until definitive deterioration of the HRQoL score as compared to the baseline score (TUDD baseline) or the best previous score (TUDD best) for different values of sample size (N), items (I), correlations between HRQoL measure (rho) and the proportion of missing data which was fixed to $\pi = 0.20$.

Table 1 - Summary of the definitions of time to quality of life score deterioration approach retained for the simulation study

MCID 5-point	Reference score	Deterioration
TTD ≥ 5-points	Baseline	Not definitive
TUDD ≥ 5-points with no further improvement 5 points as compared to reference score	Baseline	Definitive
TTD ≥ 5-points	Best previous score	Not definitive
TUDD ≥ 5-points with no further improvement 5 points as compared to reference score	Best previous score	Definitive

MCID: Minimal Clinically Important Difference
TTD: Time To Deterioration
TUDD: Time Until Definitive Deterioration

Table 1 - Type I error rate of the test of interaction between treatment arm and time for simulations with complete data

N	I	J	ρ	5 measures						10 measures					
				SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM
100	1	4	0.4	0.056	0.076	0.062	0.062	0.068	0.068	0.042	0.058	0.046	0.070	0.060	0.100
			0.7	0.048	0.028	0.058	0.052	0.060	0.048	0.032	0.050	0.066	0.042	0.040	0.044
			0.9	0.062	0.032	0.040	0.056	0.048	0.064	0.044	0.066	0.054	0.056	0.076	0.046
	2	4	0.4	0.062	0.058	0.048	0.056	0.064	0.056	0.054	0.046	0.042	0.046	0.060	0.050
			0.7	0.048	0.064	0.064	0.056	0.062	0.048	0.048	0.058	0.048	0.052	0.064	0.044
			0.9	0.054	0.054	0.054	0.050	0.044	0.066	0.040	0.066	0.058	0.056	0.056	0.036
	4	4	0.4	0.080	0.062	0.070	0.064	0.072	0.070	0.052	0.048	0.060	0.070	0.064	0.046
			0.7	0.078	0.040	0.048	0.052	0.068	0.072	0.048	0.060	0.060	0.050	0.064	0.052
			0.9	0.078	0.046	0.046	0.056	0.060	0.070	0.042	0.048	0.056	0.068	0.048	0.046
200	1	4	0.4	0.038	0.050	0.046	0.046	0.040	0.054	0.068	0.052	0.048	0.072	0.050	0.138
			0.7	0.056	0.064	0.042	0.054	0.062	0.054	0.054	0.052	0.064	0.054	0.054	0.058
			0.9	0.040	0.034	0.052	0.058	0.048	0.034	0.034	0.038	0.036	0.042	0.042	0.046
	2	4	0.4	0.046	0.060	0.048	0.046	0.042	0.046	0.044	0.064	0.064	0.068	0.054	0.048
			0.7	0.054	0.072	0.072	0.074	0.060	0.054	0.032	0.054	0.050	0.048	0.044	0.046
			0.9	0.068	0.054	0.060	0.046	0.046	0.066	0.040	0.044	0.058	0.058	0.044	0.042
		7	0.4	0.040	0.036	0.046	0.034	0.046	0.040	0.050	0.056	0.040	0.062	0.054	0.054
			0.7	0.044	0.046	0.050	0.056	0.056	0.040	0.038	0.066	0.058	0.056	0.054	0.042
			0.9	0.062	0.046	0.052	0.054	0.042	0.056	0.042	0.058	0.040	0.060	0.054	0.040
	4	4	0.4	0.042	0.056	0.050	0.052	0.060	0.036	0.048	0.032	0.056	0.052	0.046	0.042
			0.7	0.044	0.050	0.060	0.052	0.052	0.052	0.038	0.052	0.058	0.058	0.066	0.056
			0.9	0.054	0.042	0.050	0.052	0.042	0.050	0.030	0.054	0.048	0.064	0.054	0.042
300	1	4	0.4	0.046	0.054	0.064	0.052	0.072	0.058	0.050	0.052	0.046	0.050	0.058	0.108
			0.7	0.076	0.058	0.066	0.068	0.066	0.074	0.034	0.042	0.044	0.054	0.054	0.046
			0.9	0.038	0.050	0.040	0.050	0.052	0.034	0.034	0.058	0.072	0.074	0.076	0.036
	2	4	0.4	0.046	0.072	0.058	0.082	0.058	0.054	0.034	0.056	0.054	0.072	0.054	0.062
			0.7	0.044	0.058	0.046	0.044	0.052	0.046	0.040	0.054	0.050	0.052	0.058	0.038
			0.9	0.052	0.060	0.064	0.050	0.046	0.054	0.050	0.064	0.062	0.038	0.050	0.044
	4	4	0.4	0.040	0.062	0.068	0.042	0.044	0.064	0.038	0.042	0.060	0.048	0.062	0.040
			0.7	0.050	0.054	0.062	0.054	0.050	0.060	0.044	0.050	0.060	0.048	0.056	0.048
			0.9	0.044	0.066	0.062	0.044	0.052	0.038	0.036	0.044	0.048	0.028	0.044	0.048

The methods compared are the Score and Mixed Model (SM), Longitudinal Partial Credit Model (LPCM), Time to HRQoL score deterioration as compared to the baseline score (TTD baseline) or the best previous score (TTD best) and time until definitive deterioration of the HRQoL score as compared to the baseline score (TUDD baseline) or the best previous score (TUDD best) for different values of sample size (N), items (I), response category per item (J) and correlations between HRQoL measure (ρ).

Table 3 - Power of the test of interaction between treatment arm and time for simulations with complete data

N	I	J	ρ	5 measurement times						10 measurement times					
				SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM
100	1	4	0.4	0.518	0.118	0.102	0.130	0.100	0.472	0.784	0.134	0.132	0.120	0.058	0.654
			0.7	0.418	0.148	0.104	0.156	0.126	0.388	0.578	0.114	0.110	0.118	0.082	0.484
			0.9	0.404	0.148	0.124	0.174	0.142	0.380	0.414	0.124	0.126	0.130	0.084	0.352
	2	4	0.4	0.590	0.182	0.146	0.218	0.138	0.528	0.874	0.172	0.158	0.192	0.080	0.726
			0.7	0.488	0.214	0.162	0.208	0.162	0.432	0.628	0.188	0.180	0.192	0.096	0.354
			0.9	0.414	0.226	0.150	0.216	0.134	0.394	0.466	0.206	0.164	0.210	0.096	0.290
	4	4	0.4	0.680	0.260	0.168	0.204	0.128	0.412	0.916	0.236	0.190	0.246	0.090	0.396
			0.7	0.550	0.290	0.214	0.246	0.148	0.394	0.664	0.282	0.228	0.258	0.116	0.214
			0.9	0.496	0.394	0.290	0.364	0.210	0.428	0.428	0.400	0.284	0.314	0.150	0.228
200	1	4	0.4	0.812	0.212	0.162	0.240	0.142	0.778	0.970	0.140	0.168	0.182	0.060	0.858
			0.7	0.644	0.208	0.168	0.204	0.148	0.622	0.872	0.170	0.184	0.216	0.066	0.798
			0.9	0.644	0.234	0.176	0.256	0.160	0.612	0.670	0.194	0.152	0.204	0.082	0.572
	2	4	0.4	0.894	0.296	0.232	0.304	0.148	0.830	0.992	0.272	0.246	0.328	0.116	0.952
			0.7	0.760	0.350	0.262	0.344	0.214	0.678	0.934	0.326	0.242	0.334	0.114	0.618
			0.9	0.720	0.416	0.324	0.440	0.276	0.710	0.726	0.398	0.316	0.400	0.156	0.494
		7	0.4	0.936	0.346	0.270	0.328	0.214	0.832	0.999	0.362	0.324	0.366	0.096	0.900
			0.7	0.826	0.468	0.392	0.458	0.284	0.678	0.926	0.444	0.358	0.360	0.124	0.458
			0.9	0.810	0.580	0.442	0.614	0.400	0.788	0.726	0.600	0.506	0.536	0.218	0.458
	4	4	0.4	0.954	0.402	0.292	0.364	0.156	0.722	0.996	0.370	0.288	0.366	0.118	0.672
			0.7	0.812	0.552	0.394	0.456	0.236	0.606	0.934	0.484	0.362	0.380	0.142	0.368
			0.9	0.796	0.678	0.518	0.632	0.388	0.760	0.728	0.608	0.510	0.494	0.196	0.374
300	1	4	0.4	0.928	0.218	0.180	0.274	0.160	0.916	0.998	0.202	0.236	0.234	0.062	0.920
			0.7	0.842	0.274	0.234	0.282	0.162	0.842	0.974	0.232	0.220	0.242	0.082	0.920
			0.9	0.820	0.338	0.248	0.386	0.234	0.796	0.856	0.244	0.236	0.300	0.108	0.772
	2	4	0.4	0.980	0.410	0.336	0.446	0.232	0.948	0.998	0.390	0.342	0.424	0.092	0.988
			0.7	0.918	0.494	0.366	0.514	0.304	0.858	0.984	0.500	0.378	0.462	0.124	0.788
			0.9	0.902	0.560	0.418	0.594	0.366	0.886	0.902	0.508	0.388	0.476	0.154	0.670
	4	4	0.4	0.990	0.578	0.416	0.510	0.242	0.856	0.998	0.560	0.398	0.486	0.172	0.822
			0.7	0.956	0.636	0.488	0.564	0.302	0.792	0.990	0.678	0.526	0.582	0.194	0.516
			0.9	0.966	0.820	0.678	0.790	0.484	0.916	0.912	0.834	0.650	0.708	0.300	0.530

The methods compared are the Score and Mixed Model (SM), Longitudinal Partial Credit Model (LPCM), Time to HRQoL score deterioration as compared to the baseline score (TTD baseline) or the best previous score (TTD best) and time until definitive deterioration of the HRQoL score as compared to the baseline score (TUDD baseline) or the best previous score (TUDD best) for different values of sample size (N), items (I), response category per item (J) and correlations between HRQoL measure (ρ).

Table 4 - Type I error of the test of interaction between treatment arm and time for datasets simulated with intermittent missing forms and monotone missing data

					5 measures						10 measures					
N	I	J	ρ	π	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM
100	1	4	0.4	0.10	0.058	0.074	0.064	0.070	0.068	0.060	0.048	0.068	0.100	0.086	0.064	0.130
				0.20	0.076	0.052	0.058	0.054	0.048	0.064	0.064	0.060	0.048	0.054	0.052	0.126
				0.30	0.068	0.070	0.068	0.062	0.066	0.066	0.052	0.056	0.056	0.048	0.054	0.082
			0.7	0.10	0.068	0.052	0.058	0.058	0.056	0.064	0.052	0.068	0.058	0.066	0.044	0.060
				0.20	0.068	0.042	0.048	0.060	0.078	0.060	0.054	0.048	0.036	0.078	0.050	0.056
				0.30	0.074	0.054	0.050	0.074	0.068	0.064	0.062	0.060	0.050	0.056	0.054	0.056
			0.9	0.10	0.078	0.048	0.042	0.054	0.052	0.082	0.040	0.052	0.048	0.042	0.064	0.036
				0.20	0.056	0.050	0.038	0.052	0.044	0.046	0.066	0.042	0.046	0.054	0.042	0.074
				0.30	0.044	0.038	0.052	0.044	0.062	0.038	0.060	0.048	0.046	0.040	0.052	0.052
	2	4	0.4	0.10	0.060	0.076	0.070	0.064	0.066	0.066	0.042	0.048	0.070	0.060	0.052	0.054
				0.20	0.064	0.040	0.048	0.058	0.064	0.066	0.042	0.068	0.048	0.060	0.056	0.042
				0.30	0.054	0.044	0.038	0.056	0.058	0.066	0.058	0.054	0.076	0.056	0.060	0.056
			0.7	0.10	0.068	0.052	0.058	0.050	0.050	0.058	0.038	0.048	0.040	0.040	0.052	0.048
				0.20	0.070	0.050	0.054	0.060	0.066	0.068	0.052	0.060	0.060	0.056	0.060	0.052
				0.30	0.072	0.058	0.062	0.066	0.076	0.070	0.068	0.068	0.056	0.056	0.054	0.052
			0.9	0.10	0.055	0.056	0.064	0.056	0.058	0.050	0.070	0.060	0.058	0.062	0.058	0.052
				0.20	0.066	0.058	0.048	0.052	0.054	0.060	0.050	0.042	0.054	0.040	0.054	0.056
				0.30	0.064	0.046	0.070	0.062	0.072	0.066	0.060	0.052	0.044	0.050	0.052	0.046
	4	4	0.4	0.10	0.074	0.068	0.076	0.070	0.060	0.056	0.064	0.050	0.058	0.056	0.050	0.050
				0.20	0.074	0.050	0.052	0.078	0.052	0.058	0.048	0.060	0.058	0.058	0.058	0.030
				0.30	0.066	0.060	0.062	0.066	0.086	0.066	0.064	0.076	0.068	0.05	0.058	0.058
			0.7	0.10	0.076	0.036	0.048	0.068	0.062	0.078	0.046	0.054	0.072	0.048	0.048	0.048
				0.20	0.074	0.036	0.052	0.066	0.068	0.080	0.076	0.054	0.044	0.070	0.052	0.052
				0.30	0.058	0.058	0.066	0.066	0.054	0.050	0.070	0.058	0.056	0.058	0.046	0.074
			0.9	0.10	0.050	0.060	0.082	0.046	0.062	0.046	0.046	0.066	0.054	0.048	0.050	0.062
				0.20	0.058	0.070	0.052	0.054	0.060	0.066	0.046	0.042	0.058	0.056	0.052	0.052
				0.30	0.058	0.062	0.042	0.066	0.066	0.058	0.034	0.034	0.054	0.064	0.072	0.068
200	1	4	0.4	0.10	0.042	0.042	0.046	0.056	0.06	0.046	0.058	0.068	0.068	0.062	0.074	0.156
				0.20	0.056	0.052	0.058	0.046	0.044	0.060	0.038	0.044	0.046	0.054	0.042	0.122
				0.30	0.064	0.072	0.078	0.052	0.068	0.054	0.046	0.062	0.050	0.080	0.060	0.102
			0.7	0.10	0.058	0.042	0.038	0.054	0.052	0.058	0.042	0.066	0.058	0.068	0.068	0.056
				0.20	0.076	0.050	0.056	0.046	0.042	0.062	0.052	0.042	0.034	0.052	0.054	0.052
				0.30	0.048	0.052	0.064	0.062	0.046	0.046	0.064	0.058	0.038	0.050	0.044	0.066
			0.9	0.10	0.050	0.072	0.06	0.052	0.054	0.050	0.032	0.054	0.036	0.060	0.052	0.036
				0.20	0.044	0.032	0.044	0.038	0.052	0.034	0.062	0.064	0.060	0.050	0.056	0.058
				0.30	0.062	0.054	0.056	0.036	0.054	0.056	0.028	0.042	0.058	0.050	0.044	0.030
	2	4	0.4	0.10	0.042	0.076	0.080	0.048	0.066	0.044	0.046	0.040	0.050	0.054	0.054	0.036
				0.20	0.048	0.060	0.070	0.056	0.056	0.052	0.046	0.034	0.056	0.056	0.064	0.058
				0.30	0.032	0.048	0.036	0.050	0.056	0.042	0.052	0.062	0.052	0.050	0.050	0.048
			0.7	0.10	0.046	0.044	0.048	0.038	0.046	0.042	0.044	0.054	0.046	0.052	0.058	0.058
				0.20	0.050	0.056	0.068	0.062	0.058	0.044	0.052	0.054	0.036	0.062	0.044	0.058
				0.30	0.062	0.048	0.040	0.066	0.072	0.060	0.038	0.044	0.038	0.038	0.040	0.044
			0.9	0.10	0.050	0.042	0.048	0.052	0.032	0.052	0.046	0.050	0.060	0.054	0.054	0.046
				0.20	0.058	0.050	0.050	0.058	0.048	0.060	0.036	0.058	0.052	0.048	0.050	0.040
				0.30	0.054	0.060	0.046	0.052	0.044	0.060	0.056	0.056	0.048	0.068	0.054	0.050

Table 4 - Type I error of the test of interaction between treatment arm and time for datasets simulated with intermittent missing forms and monotone missing data *(continued)*

					5 measures						10 measures					
N	I	J	ρ	π	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM
200	2	7	0.4	0.10	0.052	0.056	0.072	0.05	0.044	0.054	0.046	0.062	0.040	0.060	0.054	0.046
				0.20	0.052	0.060	0.064	0.044	0.064	0.052	0.040	0.048	0.022	0.044	0.050	0.056
				0.30	0.046	0.068	0.08	0.058	0.062	0.048	0.046	0.056	0.048	0.054	0.066	0.058
			0.7	0.10	0.044	0.056	0.054	0.056	0.044	0.050	0.036	0.052	0.062	0.050	0.048	0.044
				0.20	0.058	0.056	0.044	0.070	0.056	0.046	0.046	0.058	0.068	0.054	0.048	0.058
				0.30	0.072	0.058	0.060	0.060	0.076	0.062	0.034	0.074	0.058	0.062	0.066	0.046
			0.9	0.10	0.054	0.066	0.066	0.044	0.042	0.060	0.044	0.040	0.040	0.044	0.036	0.058
				0.20	0.048	0.072	0.048	0.048	0.070	0.048	0.048	0.062	0.050	0.046	0.046	0.056
				0.30	0.060	0.072	0.054	0.082	0.072	0.068	0.056	0.052	0.044	0.066	0.062	0.044
	4	4	0.4	0.10	0.048	0.068	0.068	0.056	0.06	0.056	0.054	0.058	0.060	0.052	0.06	0.056
				0.20	0.050	0.062	0.052	0.064	0.052	0.052	0.036	0.044	0.058	0.048	0.038	0.040
				0.30	0.070	0.060	0.048	0.062	0.060	0.064	0.058	0.05	0.044	0.062	0.052	0.050
			0.7	0.10	0.040	0.040	0.056	0.042	0.044	0.052	0.042	0.056	0.062	0.054	0.056	0.052
				0.20	0.040	0.052	0.052	0.064	0.058	0.054	0.050	0.060	0.036	0.050	0.056	0.062
				0.30	0.054	0.054	0.054	0.056	0.064	0.052	0.042	0.046	0.056	0.044	0.052	0.054
			0.9	0.10	0.060	0.062	0.058	0.06	0.054	0.052	0.038	0.054	0.048	0.048	0.030	0.040
				0.20	0.068	0.044	0.048	0.048	0.050	0.064	0.038	0.052	0.054	0.050	0.066	0.038
				0.30	0.054	0.046	0.050	0.056	0.054	0.044	0.052	0.058	0.068	0.042	0.050	0.048
300	1	4	0.4	0.10	0.052	0.044	0.048	0.054	0.052	0.060	0.050	0.050	0.042	0.054	0.036	0.110
				0.20	0.052	0.050	0.036	0.048	0.040	0.052	0.030	0.042	0.052	0.054	0.058	0.092
				0.30	0.040	0.058	0.054	0.066	0.072	0.042	0.056	0.040	0.036	0.030	0.046	0.104
			0.7	0.10	0.032	0.054	0.048	0.046	0.038	0.034	0.040	0.066	0.070	0.080	0.060	0.038
				0.20	0.042	0.072	0.058	0.056	0.044	0.044	0.042	0.054	0.046	0.060	0.038	0.050
				0.30	0.046	0.050	0.048	0.048	0.046	0.042	0.046	0.052	0.044	0.064	0.052	0.040
			0.9	0.10	0.048	0.058	0.040	0.044	0.036	0.052	0.056	0.048	0.046	0.044	0.074	0.058
				0.20	0.050	0.044	0.048	0.032	0.038	0.042	0.048	0.048	0.038	0.050	0.028	0.056
				0.30	0.054	0.040	0.050	0.040	0.040	0.048	0.048	0.046	0.048	0.040	0.052	0.048
	2	4	0.4	0.10	0.042	0.052	0.050	0.058	0.058	0.050	0.038	0.056	0.058	0.052	0.056	0.042
				0.20	0.046	0.040	0.046	0.060	0.050	0.048	0.058	0.060	0.038	0.042	0.042	0.062
				0.30	0.038	0.036	0.046	0.046	0.044	0.042	0.050	0.056	0.058	0.068	0.056	0.050
			0.7	0.10	0.038	0.044	0.042	0.048	0.040	0.042	0.040	0.042	0.050	0.048	0.048	0.046
				0.20	0.052	0.048	0.056	0.042	0.048	0.054	0.044	0.062	0.056	0.044	0.044	0.048
				0.30	0.036	0.052	0.050	0.046	0.044	0.048	0.034	0.036	0.050	0.054	0.064	0.034
			0.9	0.10	0.074	0.062	0.042	0.064	0.058	0.072	0.044	0.054	0.066	0.056	0.058	0.050
				0.20	0.042	0.060	0.052	0.046	0.044	0.052	0.042	0.048	0.066	0.038	0.040	0.030
				0.30	0.066	0.034	0.044	0.034	0.042	0.074	0.042	0.034	0.040	0.024	0.046	0.030
	4	4	0.4	0.10	0.042	0.058	0.058	0.066	0.056	0.058	0.046	0.066	0.042	0.050	0.058	0.056
				0.20	0.050	0.050	0.056	0.064	0.042	0.054	0.036	0.052	0.048	0.056	0.046	0.040
				0.30	0.050	0.060	0.034	0.044	0.038	0.038	0.038	0.042	0.036	0.068	0.05	0.048
			0.7	0.10	0.034	0.054	0.044	0.048	0.048	0.038	0.034	0.056	0.048	0.044	0.056	0.038
				0.20	0.038	0.054	0.056	0.064	0.054	0.036	0.034	0.042	0.036	0.048	0.050	0.040
				0.30	0.038	0.036	0.038	0.050	0.046	0.040	0.044	0.048	0.038	0.052	0.060	0.05
			0.9	0.10	0.038	0.048	0.048	0.044	0.060	0.048	0.038	0.050	0.032	0.046	0.048	0.036
				0.20	0.044	0.054	0.042	0.038	0.038	0.040	0.040	0.036	0.028	0.040	0.062	0.048
				0.30	0.048	0.058	0.048	0.054	0.050	0.044	0.038	0.046	0.048	0.052	0.050	0.036

The methods compared are Score and Mixed Model (SM), longitudinal Partial Credit Model (LPCM), Time to HRQoL score deterioration as compared to the baseline score (TTD baseline) or the best previous score (TTD best) and time until definitive deterioration of the HRQoL score as compared to the baseline score (TUDD baseline) or the best previous score (TUDD best) for different values of sample size (N), items (I), response category per item (J), correlations between HRQoL measure (ρ) and proportion of missing data (π).

Table 5 - Power of the test of interaction between treatment arm and time for datasets simulated with intermittent missing forms and monotone missing data

					5 measures						10 measures					
N	I	J	ρ	π	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM
100	1	4	0.4	0.10	0.486	0.136	0.118	0.144	0.094	0.448	0.732	0.086	0.112	0.110	0.052	0.614
				0.20	0.434	0.120	0.098	0.134	0.090	0.376	0.666	0.100	0.104	0.136	0.094	0.610
				0.30	0.356	0.112	0.100	0.108	0.078	0.306	0.610	0.114	0.104	0.124	0.076	0.528
			0.7	0.10	0.396	0.124	0.100	0.140	0.102	0.366	0.568	0.118	0.130	0.108	0.094	0.502
				0.20	0.358	0.098	0.088	0.134	0.104	0.336	0.534	0.100	0.094	0.128	0.058	0.478
				0.30	0.342	0.126	0.092	0.120	0.088	0.318	0.490	0.134	0.098	0.146	0.094	0.430
			0.9	0.10	0.372	0.134	0.122	0.118	0.114	0.336	0.374	0.128	0.102	0.146	0.086	0.334
				0.20	0.366	0.124	0.136	0.138	0.134	0.342	0.394	0.148	0.098	0.146	0.070	0.352
				0.30	0.338	0.138	0.112	0.150	0.090	0.308	0.384	0.136	0.108	0.138	0.090	0.340
	2	4	0.4	0.10	0.536	0.170	0.146	0.190	0.120	0.492	0.802	0.150	0.134	0.196	0.076	0.684
				0.20	0.516	0.176	0.110	0.190	0.130	0.452	0.766	0.156	0.150	0.154	0.080	0.647
				0.30	0.502	0.168	0.132	0.200	0.164	0.452	0.712	0.178	0.140	0.168	0.090	0.606
			0.7	0.10	0.474	0.224	0.168	0.208	0.144	0.414	0.596	0.170	0.150	0.182	0.054	0.380
				0.20	0.446	0.206	0.164	0.192	0.136	0.402	0.560	0.182	0.136	0.200	0.086	0.400
				0.30	0.404	0.160	0.142	0.162	0.132	0.370	0.506	0.148	0.126	0.192	0.094	0.376
			0.9	0.10	0.438	0.246	0.198	0.270	0.174	0.428	0.436	0.206	0.164	0.202	0.102	0.312
				0.20	0.438	0.218	0.168	0.234	0.172	0.424	0.424	0.218	0.162	0.218	0.118	0.326
				0.30	0.432	0.200	0.140	0.216	0.164	0.402	0.462	0.226	0.166	0.220	0.126	0.372
	4	4	0.4	0.10	0.632	0.212	0.160	0.204	0.142	0.422	0.856	0.216	0.172	0.212	0.112	0.404
				0.20	0.622	0.182	0.150	0.206	0.142	0.418	0.828	0.224	0.162	0.220	0.094	0.440
				0.30	0.548	0.184	0.146	0.206	0.140	0.404	0.760	0.178	0.122	0.208	0.104	0.410
			0.7	0.10	0.520	0.266	0.214	0.252	0.172	0.380	0.652	0.264	0.210	0.220	0.124	0.258
				0.20	0.514	0.256	0.190	0.248	0.164	0.406	0.626	0.274	0.188	0.262	0.132	0.286
				0.30	0.462	0.232	0.174	0.224	0.160	0.344	0.580	0.216	0.164	0.210	0.116	0.306
			0.9	0.10	0.524	0.368	0.288	0.368	0.250	0.474	0.466	0.340	0.294	0.296	0.140	0.292
				0.20	0.534	0.298	0.216	0.338	0.212	0.486	0.470	0.320	0.228	0.308	0.150	0.314
				0.30	0.516	0.312	0.220	0.314	0.206	0.500	0.478	0.292	0.208	0.290	0.156	0.328
200	1	4	0.4	0.10	0.768	0.180	0.140	0.194	0.138	0.734	0.946	0.130	0.136	0.172	0.072	0.844
				0.20	0.742	0.166	0.158	0.200	0.146	0.698	0.918	0.150	0.118	0.184	0.078	0.868
				0.30	0.684	0.172	0.146	0.212	0.136	0.662	0.862	0.182	0.136	0.198	0.090	0.836
			0.7	0.10	0.652	0.208	0.160	0.202	0.162	0.630	0.838	0.146	0.168	0.174	0.076	0.774
				0.20	0.644	0.206	0.178	0.218	0.162	0.626	0.810	0.212	0.190	0.216	0.114	0.740
				0.30	0.634	0.222	0.136	0.208	0.146	0.616	0.764	0.156	0.142	0.202	0.108	0.716
			0.9	0.10	0.638	0.212	0.144	0.262	0.184	0.626	0.634	0.196	0.162	0.212	0.104	0.552
				0.20	0.624	0.252	0.182	0.266	0.184	0.598	0.644	0.188	0.160	0.232	0.134	0.590
				0.30	0.562	0.204	0.180	0.232	0.190	0.550	0.638	0.202	0.166	0.242	0.114	0.582
	2	4	0.4	0.10	0.852	0.290	0.232	0.290	0.170	0.792	0.990	0.292	0.234	0.322	0.114	0.926
				0.20	0.822	0.284	0.178	0.282	0.200	0.782	0.974	0.262	0.210	0.270	0.120	0.906
				0.30	0.760	0.256	0.196	0.280	0.184	0.740	0.948	0.262	0.170	0.282	0.116	0.888
			0.7	0.10	0.754	0.318	0.236	0.332	0.186	0.690	0.902	0.316	0.244	0.280	0.082	0.650
				0.20	0.748	0.346	0.226	0.378	0.218	0.692	0.850	0.352	0.274	0.322	0.116	0.632
				0.30	0.682	0.280	0.212	0.320	0.212	0.644	0.818	0.282	0.204	0.292	0.146	0.678
			0.9	0.10	0.720	0.442	0.324	0.442	0.280	0.714	0.720	0.376	0.288	0.380	0.164	0.546
				0.20	0.694	0.368	0.264	0.402	0.248	0.676	0.684	0.394	0.290	0.398	0.160	0.574
				0.30	0.718	0.366	0.228	0.412	0.268	0.716	0.702	0.356	0.262	0.406	0.194	0.614

Table 5 - Power of the test of interaction between treatment arm and time for datasets simulated with intermittent missing forms and monotone missing data *(continued)*

					5 measures						10 measures					
N	I	J	ρ	π	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM
200	2	7	0.4	0.10	0.898	0.346	0.270	0.370	0.242	0.814	0.998	0.322	0.254	0.362	0.126	0.868
				0.20	0.896	0.326	0.226	0.352	0.196	0.776	0.986	0.330	0.238	0.338	0.128	0.878
				0.30	0.852	0.336	0.220	0.316	0.206	0.770	0.972	0.356	0.186	0.350	0.166	0.860
			0.7	0.10	0.816	0.452	0.348	0.456	0.294	0.72	0.92	0.472	0.338	0.412	0.124	0.478
				0.20	0.784	0.448	0.328	0.416	0.278	0.666	0.924	0.424	0.308	0.408	0.154	0.554
				0.30	0.786	0.394	0.254	0.404	0.260	0.71	0.862	0.438	0.29	0.374	0.170	0.616
			0.9	0.10	0.794	0.572	0.438	0.618	0.442	0.782	0.726	0.612	0.414	0.520	0.208	0.478
				0.20	0.788	0.560	0.418	0.616	0.402	0.782	0.766	0.594	0.416	0.528	0.23	0.566
				0.30	0.790	0.522	0.394	0.560	0.418	0.766	0.744	0.534	0.358	0.524	0.272	0.600
	4	4	0.4	0.10	0.920	0.360	0.246	0.380	0.214	0.706	0.996	0.388	0.316	0.362	0.142	0.682
				0.20	0.880	0.382	0.262	0.352	0.212	0.684	0.984	0.378	0.228	0.330	0.122	0.702
				0.30	0.844	0.336	0.204	0.340	0.230	0.694	0.972	0.350	0.194	0.368	0.164	0.758
			0.7	0.10	0.812	0.470	0.334	0.412	0.248	0.652	0.910	0.462	0.366	0.392	0.156	0.434
				0.20	0.816	0.468	0.316	0.402	0.216	0.686	0.890	0.442	0.314	0.386	0.174	0.482
				0.30	0.762	0.396	0.284	0.390	0.266	0.652	0.862	0.422	0.266	0.412	0.164	0.522
			0.9	0.10	0.818	0.620	0.434	0.596	0.400	0.776	0.750	0.608	0.438	0.530	0.232	0.442
				0.20	0.826	0.548	0.412	0.632	0.442	0.812	0.738	0.566	0.364	0.520	0.230	0.532
				0.30	0.808	0.520	0.360	0.536	0.364	0.800	0.744	0.570	0.368	0.508	0.246	0.620
300	1	4	0.4	0.10	0.908	0.234	0.204	0.300	0.212	0.880	0.996	0.218	0.228	0.252	0.088	0.934
				0.20	0.862	0.238	0.178	0.282	0.186	0.842	0.980	0.220	0.196	0.250	0.084	0.964
				0.30	0.852	0.240	0.194	0.260	0.192	0.820	0.962	0.216	0.178	0.236	0.090	0.946
			0.7	0.10	0.842	0.320	0.254	0.336	0.212	0.822	0.956	0.202	0.226	0.252	0.114	0.910
				0.20	0.820	0.286	0.224	0.294	0.196	0.800	0.926	0.236	0.208	0.284	0.138	0.888
				0.30	0.756	0.266	0.214	0.278	0.198	0.736	0.910	0.276	0.192	0.290	0.144	0.886
			0.9	0.10	0.784	0.310	0.222	0.320	0.238	0.778	0.836	0.250	0.210	0.292	0.126	0.776
				0.20	0.762	0.276	0.208	0.314	0.208	0.758	0.834	0.266	0.184	0.318	0.140	0.798
				0.30	0.746	0.278	0.186	0.296	0.192	0.740	0.790	0.302	0.212	0.306	0.182	0.740
	2	4	0.4	0.10	0.966	0.350	0.278	0.398	0.238	0.930	0.996	0.398	0.298	0.428	0.100	0.980
				0.20	0.950	0.332	0.224	0.400	0.230	0.914	0.998	0.374	0.284	0.436	0.130	0.986
				0.30	0.938	0.354	0.272	0.372	0.248	0.896	0.994	0.370	0.214	0.398	0.128	0.980
			0.7	0.10	0.914	0.458	0.322	0.454	0.284	0.858	0.980	0.432	0.328	0.436	0.114	0.802
				0.20	0.890	0.400	0.278	0.434	0.260	0.858	0.964	0.416	0.274	0.418	0.166	0.836
				0.30	0.862	0.416	0.302	0.442	0.262	0.818	0.944	0.450	0.268	0.432	0.196	0.830
			0.9	0.10	0.892	0.564	0.386	0.582	0.380	0.880	0.880	0.578	0.390	0.528	0.194	0.716
				0.20	0.876	0.514	0.342	0.554	0.338	0.870	0.894	0.546	0.392	0.564	0.246	0.786
				0.30	0.866	0.500	0.350	0.534	0.348	0.848	0.872	0.534	0.346	0.512	0.242	0.778
	4	4	0.4	0.10	0.988	0.508	0.36	0.504	0.262	0.880	0.998	0.514	0.394	0.506	0.172	0.876
				0.20	0.972	0.454	0.302	0.466	0.278	0.856	0.998	0.492	0.344	0.500	0.184	0.882
				0.30	0.968	0.428	0.298	0.472	0.306	0.836	0.996	0.488	0.290	0.510	0.194	0.874
			0.7	0.10	0.964	0.660	0.456	0.618	0.324	0.846	0.992	0.636	0.480	0.538	0.200	0.596
				0.20	0.932	0.626	0.430	0.580	0.338	0.834	0.980	0.606	0.400	0.560	0.236	0.646
				0.30	0.934	0.574	0.386	0.572	0.364	0.856	0.972	0.598	0.376	0.562	0.246	0.724
			0.9	0.10	0.944	0.786	0.586	0.794	0.550	0.938	0.898	0.780	0.626	0.714	0.316	0.632
				0.20	0.914	0.714	0.556	0.780	0.562	0.896	0.910	0.774	0.564	0.706	0.350	0.692
				0.30	0.948	0.678	0.506	0.726	0.520	0.936	0.892	0.758	0.504	0.692	0.386	0.772

Methods compared are the Score and Mixed Model (SM), longitudinal Partial Credit Model (LPCM), Time to HRQoL score deterioration as compared to the baseline score (TTD baseline) or the best previous score (TTD best) and time until definitive deterioration of the HRQoL score as compared to the baseline score (TUDD baseline) or the best previous score (TUDD best) for different values of sample size (N), items (I), response category per item (J), correlations between HRQoL measure (ρ) and proportion of missing data (π).

Additional files

Additional file 1 – Complementary results obtained with intermittent missing items and monotone missing data

This file contains two additional tables A1 and A2 with respectively the type I error rate and statistical power of the test of interaction between treatment arm and time for datasets simulated with intermittent missing items and monotone missing data.

Table A1 - Type I error rate of the test of interaction between treatment arm and time for datasets simulated with intermittent missing items and monotone missing data

					5 measures						10 measures					
N	I	J	ρ	π	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM
100	2	4	0.4	0.10	0.054	0.058	0.084	0.066	0.076	0.064	0.042	0.046	0.040	0.056	0.056	0.052
				0.20	0.058	0.058	0.058	0.062	0.064	0.062	0.056	0.052	0.056	0.052	0.060	0.064
				0.30	0.070	0.068	0.052	0.070	0.054	0.066	0.060	0.048	0.054	0.058	0.046	0.064
			0.7	0.10	0.054	0.038	0.046	0.064	0.054	0.060	0.044	0.042	0.036	0.046	0.048	0.044
				0.20	0.060	0.038	0.052	0.058	0.078	0.066	0.060	0.040	0.054	0.056	0.066	0.044
				0.30	0.062	0.042	0.042	0.052	0.046	0.066	0.046	0.060	0.054	0.06	0.056	0.034
			0.9	0.10	0.050	0.050	0.054	0.046	0.060	0.052	0.056	0.048	0.054	0.052	0.072	0.052
				0.20	0.052	0.040	0.040	0.056	0.068	0.058	0.064	0.040	0.048	0.058	0.050	0.044
				0.30	0.042	0.044	0.048	0.044	0.064	0.048	0.068	0.050	0.050	0.054	0.066	0.072
	4	4	0.4	0.10	0.050	0.056	0.064	0.066	0.058	0.054	0.068	0.078	0.072	0.058	0.056	0.046
				0.20	0.052	0.064	0.070	0.066	0.064	0.066	0.048	0.066	0.050	0.068	0.048	0.048
				0.30	0.060	0.060	0.058	0.062	0.060	0.086	0.058	0.060	0.046	0.072	0.074	0.048
			0.7	0.10	0.060	0.054	0.056	0.060	0.050	0.068	0.056	0.060	0.066	0.060	0.06	0.056
				0.20	0.064	0.058	0.046	0.056	0.080	0.068	0.042	0.070	0.058	0.042	0.050	0.056
				0.30	0.064	0.042	0.044	0.066	0.064	0.064	0.052	0.046	0.054	0.054	0.060	0.046
			0.9	0.10	0.052	0.058	0.076	0.076	0.050	0.060	0.064	0.032	0.056	0.062	0.066	0.046
				0.20	0.038	0.046	0.048	0.044	0.044	0.036	0.064	0.058	0.058	0.046	0.076	0.052
				0.30	0.082	0.062	0.074	0.066	0.072	0.068	0.050	0.056	0.048	0.058	0.050	0.058
200	2	4	0.4	0.10	0.048	0.058	0.044	0.042	0.050	0.044	0.048	0.046	0.062	0.056	0.042	0.044
				0.20	0.044	0.064	0.064	0.054	0.050	0.044	0.048	0.058	0.044	0.042	0.044	0.044
				0.30	0.046	0.064	0.058	0.060	0.062	0.050	0.064	0.066	0.068	0.068	0.058	0.064
			0.7	0.10	0.062	0.060	0.078	0.052	0.054	0.052	0.040	0.064	0.060	0.052	0.05	0.036
				0.20	0.054	0.056	0.068	0.056	0.072	0.056	0.044	0.042	0.034	0.046	0.066	0.046
				0.30	0.088	0.056	0.062	0.056	0.052	0.080	0.044	0.044	0.056	0.058	0.066	0.058
			0.9	0.10	0.052	0.052	0.056	0.056	0.046	0.056	0.046	0.048	0.064	0.046	0.06	0.048
				0.20	0.060	0.056	0.042	0.058	0.054	0.064	0.042	0.044	0.050	0.058	0.058	0.032
				0.30	0.024	0.060	0.048	0.052	0.042	0.040	0.046	0.068	0.062	0.052	0.040	0.052
		7	0.4	0.10	0.060	0.056	0.046	0.066	0.048	0.058	0.056	0.076	0.070	0.052	0.058	0.046
				0.20	0.052	0.054	0.032	0.062	0.056	0.044	0.046	0.052	0.052	0.056	0.064	0.048
				0.30	0.050	0.032	0.046	0.056	0.056	0.054	0.042	0.052	0.056	0.052	0.068	0.032
			0.7	0.10	0.036	0.064	0.048	0.064	0.044	0.038	0.042	0.050	0.058	0.062	0.056	0.048
				0.20	0.060	0.060	0.072	0.046	0.046	0.058	0.036	0.066	0.052	0.058	0.066	0.032
				0.30	0.048	0.050	0.062	0.050	0.054	0.052	0.046	0.038	0.052	0.062	0.050	0.042
			0.9	0.10	0.056	0.048	0.046	0.046	0.046	0.050	0.046	0.052	0.052	0.048	0.054	0.044
				0.20	0.050	0.054	0.058	0.052	0.050	0.044	0.046	0.042	0.056	0.046	0.066	0.058
				0.30	0.05	0.058	0.048	0.042	0.052	0.058	0.052	0.038	0.054	0.068	0.050	0.052

Table A1 - Type I error rate of the test of interaction between treatment arm and time for datasets simulated with intermittent missing items and monotone missing data *(continued)*

					5 measures						10 measures					
N	I	J	ρ	π	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM
	4	4	0.4	0.10	0.076	0.056	0.046	0.042	0.044	0.040	0.050	0.058	0.048	0.050	0.054	0.040
				0.20	0.044	0.064	0.072	0.056	0.062	0.034	0.042	0.048	0.060	0.054	0.070	0.050
				0.30	0.050	0.054	0.048	0.050	0.036	0.040	0.048	0.052	0.046	0.062	0.042	0.042
			0.7	0.10	0.056	0.054	0.042	0.052	0.052	0.046	0.050	0.046	0.046	0.066	0.056	0.060
				0.20	0.042	0.046	0.048	0.046	0.050	0.038	0.060	0.052	0.042	0.048	0.046	0.070
				0.30	0.060	0.066	0.044	0.060	0.036	0.036	0.042	0.062	0.064	0.066	0.052	0.042
			0.9	0.10	0.058	0.054	0.056	0.056	0.058	0.062	0.030	0.038	0.048	0.042	0.044	0.036
				0.20	0.064	0.062	0.054	0.064	0.054	0.070	0.050	0.076	0.070	0.062	0.052	0.044
				0.30	0.048	0.036	0.030	0.028	0.028	0.048	0.052	0.064	0.066	0.058	0.042	0.042
300	2	4	0.4	0.10	0.042	0.046	0.050	0.054	0.052	0.048	0.040	0.056	0.042	0.062	0.052	0.046
				0.20	0.042	0.052	0.052	0.044	0.030	0.046	0.034	0.036	0.040	0.046	0.042	0.052
				0.30	0.036	0.048	0.056	0.048	0.048	0.050	0.050	0.032	0.040	0.046	0.046	0.050
			0.7	0.10	0.054	0.058	0.062	0.056	0.052	0.058	0.034	0.060	0.054	0.040	0.044	0.038
				0.20	0.038	0.044	0.044	0.058	0.060	0.042	0.046	0.060	0.050	0.058	0.064	0.046
				0.30	0.050	0.034	0.040	0.038	0.048	0.056	0.036	0.050	0.048	0.028	0.056	0.038
			0.9	0.10	0.044	0.044	0.060	0.052	0.050	0.042	0.048	0.042	0.048	0.048	0.054	0.056
				0.20	0.036	0.036	0.048	0.042	0.044	0.044	0.046	0.050	0.038	0.034	0.034	0.038
				0.30	0.040	0.048	0.050	0.036	0.040	0.046	0.046	0.048	0.046	0.054	0.052	0.058
	4	4	0.4	0.10	0.038	0.038	0.044	0.042	0.044	0.048	0.036	0.064	0.070	0.046	0.054	0.046
				0.20	0.048	0.056	0.060	0.054	0.062	0.058	0.028	0.064	0.066	0.052	0.050	0.038
				0.30	0.044	0.054	0.054	0.058	0.052	0.066	0.052	0.042	0.038	0.054	0.056	0.044
			0.7	0.10	0.040	0.054	0.058	0.050	0.044	0.046	0.054	0.038	0.050	0.048	0.060	0.042
				0.20	0.038	0.064	0.058	0.048	0.070	0.044	0.042	0.042	0.050	0.042	0.046	0.060
				0.30	0.038	0.052	0.048	0.040	0.056	0.062	0.042	0.04	0.048	0.056	0.048	0.042
			0.9	0.10	0.042	0.046	0.052	0.040	0.054	0.044	0.036	0.036	0.056	0.05	0.050	0.040
				0.20	0.034	0.038	0.060	0.050	0.050	0.022	0.040	0.038	0.060	0.050	0.056	0.036
				0.30	0.050	0.062	0.064	0.042	0.046	0.056	0.054	0.042	0.064	0.048	0.056	0.058

Methods compared are the Score and Mixed Model (SM), longitudinal Partial Credit Model (LPCM), Time to HRQoL score deterioration as compared to the baseline score (TTD baseline) or the best previous score (TTD best) and time until definitive deterioration of the HRQoL score as compared to the baseline score (TUDD baseline) or the best previous score (TUDD best) for different values of sample size (N), items (I), response category per item (J), correlations between HRQoL measure (ρ) and proportion of missing data (π).

Table A2 - Power of the test of interaction between treatment arm and time for the datasets simulated with intermittent missing items and monotone missing data

					5 measures						10 measures					
N	I	J	ρ	π	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM
100	2	4	0.4	0.10	0.582	0.180	0.142	0.178	0.116	0.524	0.840	0.176	0.158	0.170	0.104	0.672
				0.20	0.534	0.186	0.146	0.218	0.146	0.476	0.762	0.174	0.150	0.182	0.096	0.638
				0.30	0.516	0.160	0.134	0.168	0.112	0.478	0.724	0.154	0.136	0.136	0.058	0.590
			0.7	0.10	0.476	0.194	0.168	0.198	0.156	0.404	0.600	0.184	0.158	0.158	0.082	0.354
				0.20	0.446	0.182	0.128	0.188	0.124	0.394	0.578	0.200	0.178	0.178	0.064	0.368
				0.30	0.418	0.192	0.144	0.176	0.132	0.370	0.552	0.156	0.116	0.196	0.088	0.422
			0.9	0.10	0.428	0.246	0.192	0.270	0.172	0.420	0.426	0.252	0.212	0.218	0.112	0.324
				0.20	0.442	0.256	0.178	0.252	0.176	0.420	0.432	0.230	0.182	0.230	0.086	0.344
				0.30	0.434	0.256	0.200	0.256	0.194	0.412	0.412	0.218	0.180	0.232	0.124	0.346
	4	4	0.4	0.10	0.592	0.134	0.108	0.178	0.148	0.396	0.858	0.19	0.154	0.202	0.102	0.412
				0.20	0.536	0.162	0.134	0.198	0.148	0.446	0.780	0.188	0.116	0.190	0.108	0.476
				0.30	0.454	0.118	0.094	0.164	0.130	0.410	0.628	0.146	0.102	0.160	0.106	0.468
			0.7	0.10	0.472	0.228	0.156	0.238	0.160	0.356	0.618	0.212	0.166	0.230	0.140	0.234
				0.20	0.486	0.212	0.164	0.204	0.164	0.398	0.608	0.188	0.130	0.186	0.130	0.268
				0.30	0.404	0.164	0.146	0.162	0.154	0.408	0.502	0.162	0.110	0.218	0.114	0.342
			0.9	0.10	0.532	0.278	0.202	0.334	0.224	0.472	0.472	0.344	0.234	0.312	0.150	0.248
				0.20	0.502	0.210	0.166	0.244	0.172	0.470	0.474	0.240	0.144	0.272	0.144	0.302
				0.30	0.460	0.200	0.144	0.238	0.166	0.508	0.478	0.232	0.180	0.238	0.148	0.342
200	2	4	0.4	0.10	0.864	0.290	0.226	0.278	0.180	0.818	0.988	0.270	0.210	0.306	0.092	0.938
				0.20	0.826	0.274	0.232	0.288	0.178	0.744	0.982	0.288	0.214	0.272	0.082	0.924
				0.30	0.802	0.270	0.206	0.292	0.178	0.772	0.960	0.272	0.214	0.292	0.092	0.900
			0.7	0.10	0.786	0.328	0.256	0.354	0.212	0.728	0.892	0.320	0.270	0.312	0.076	0.618
				0.20	0.726	0.320	0.264	0.354	0.198	0.684	0.840	0.336	0.264	0.340	0.106	0.644
				0.30	0.696	0.312	0.214	0.318	0.162	0.650	0.824	0.372	0.282	0.286	0.138	0.670
			0.9	0.10	0.722	0.722	0.340	0.454	0.290	0.724	0.692	0.378	0.284	0.374	0.136	0.528
				0.20	0.686	0.386	0.268	0.440	0.270	0.716	0.672	0.394	0.276	0.410	0.128	0.584
				0.30	0.670	0.404	0.294	0.436	0.270	0.692	0.686	0.350	0.252	0.362	0.162	0.618
		7	0.4	0.10	0.922	0.386	0.278	0.386	0.228	0.808	0.994	0.408	0.324	0.382	0.112	0.880
				0.20	0.906	0.376	0.274	0.386	0.262	0.812	0.990	0.370	0.304	0.382	0.136	0.900
				0.30	0.854	0.350	0.288	0.340	0.232	0.778	0.970	0.364	0.292	0.382	0.108	0.888
			0.7	0.10	0.816	0.460	0.370	0.444	0.302	0.702	0.912	0.498	0.366	0.38	0.124	0.516
				0.20	0.792	0.454	0.372	0.448	0.268	0.714	0.904	0.464	0.384	0.404	0.156	0.606
				0.30	0.794	0.470	0.350	0.452	0.270	0.724	0.88	0.442	0.318	0.396	0.148	0.628
			0.9	0.10	0.784	0.594	0.484	0.594	0.440	0.78	0.740	0.614	0.510	0.526	0.188	0.510
				0.20	0.794	0.596	0.418	0.632	0.434	0.818	0.724	0.58	0.468	0.514	0.216	0.546
				0.30	0.800	0.558	0.404	0.616	0.428	0.800	0.734	0.594	0.440	0.556	0.24	0.618
	4	4	0.4	0.10	0.888	0.320	0.200	0.366	0.218	0.716	0.990	0.346	0.206	0.362	0.124	0.704
				0.20	0.840	0.240	0.148	0.264	0.180	0.698	0.972	0.294	0.148	0.324	0.158	0.748
				0.30	0.784	0.222	0.174	0.254	0.188	0.698	0.912	0.252	0.136	0.28	0.152	0.770
			0.7	0.10	0.818	0.414	0.278	0.410	0.258	0.674	0.912	0.396	0.248	0.364	0.154	0.426
				0.20	0.738	0.310	0.220	0.334	0.224	0.652	0.868	0.322	0.214	0.350	0.176	0.484
				0.30	0.716	0.282	0.202	0.280	0.210	0.712	0.800	0.266	0.142	0.316	0.180	0.540
			0.9	0.10	0.844	0.538	0.388	0.568	0.388	0.806	0.768	0.580	0.360	0.506	0.234	0.460
				0.20	0.778	0.442	0.290	0.488	0.340	0.776	0.790	0.436	0.266	0.488	0.230	0.548
				0.30	0.728	0.328	0.248	0.378	0.278	0.786	0.708	0.412	0.230	0.414	0.226	0.584

Table A2 - Power of the test of interaction between treatment arm and time for the datasets simulated with intermittent missing items and monotone missing data *(continued)*

					5 measures						10 measures					
N	I	J	ρ	π	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM	SM	TTD baseline	TTD best	TUDD baseline	TUDD best	LPCM
300	2	4	0.4	0.10	0.970	0.430	0.324	0.432	0.240	0.936	0.998	0.386	0.340	0.374	0.09	0.984
				0.20	0.956	0.374	0.276	0.394	0.250	0.924	0.996	0.456	0.332	0.444	0.096	0.982
				0.30	0.934	0.426	0.296	0.428	0.254	0.884	0.994	0.404	0.302	0.400	0.088	0.984
			0.7	0.10	0.922	0.498	0.332	0.510	0.304	0.882	0.978	0.466	0.368	0.466	0.130	0.806
				0.20	0.894	0.484	0.356	0.510	0.322	0.870	0.970	0.452	0.322	0.496	0.156	0.866
				0.30	0.866	0.458	0.376	0.466	0.310	0.818	0.956	0.440	0.314	0.424	0.158	0.866
			0.9	0.10	0.880	0.588	0.438	0.610	0.376	0.866	0.878	0.538	0.424	0.508	0.184	0.702
				0.20	0.860	0.568	0.372	0.596	0.372	0.860	0.852	0.532	0.410	0.480	0.176	0.740
				0.30	0.854	0.520	0.382	0.592	0.386	0.850	0.870	0.498	0.360	0.520	0.202	0.802
	4	4	0.4	0.10	0.978	0.438	0.278	0.474	0.266	0.894	0.998	0.476	0.26	0.49	0.186	0.856
				0.20	0.950	0.372	0.234	0.394	0.250	0.876	0.992	0.376	0.190	0.446	0.212	0.876
				0.30	0.902	0.318	0.244	0.346	0.250	0.878	0.978	0.336	0.192	0.422	0.234	0.900
			0.7	0.10	0.934	0.560	0.394	0.562	0.338	0.828	0.990	0.584	0.364	0.536	0.226	0.604
				0.20	0.918	0.476	0.320	0.514	0.320	0.850	0.966	0.476	0.278	0.482	0.266	0.646
				0.30	0.836	0.354	0.284	0.370	0.268	0.840	0.928	0.430	0.252	0.45	0.238	0.732
			0.9	0.10	0.944	0.686	0.490	0.736	0.526	0.918	0.916	0.720	0.466	0.704	0.340	0.628
				0.20	0.926	0.564	0.412	0.592	0.414	0.922	0.916	0.636	0.364	0.678	0.372	0.758
				0.30	0.880	0.474	0.344	0.538	0.398	0.908	0.904	0.516	0.308	0.544	0.352	0.782

Methods compared are the Score and Mixed Model (SM), longitudinal Partial Credit Model (LPCM), Time to HRQoL score deterioration as compared to the baseline score (TTD baseline) or the best previous score (TTD best) and time until definitive deterioration of the HRQoL score as compared to the baseline score (TUDD baseline) or the best previous score (TUDD best) for different values of sample size (N), items (I), response category per item (J), correlations between HRQoL measure (ρ) and proportion of missing data (π)

4. Caractérisation de l'occurrence de l'effet Response Shift

4.1. Analyses factorielles et modèles issus de la théorie de réponse à l'item

Résumé

Introduction

En oncologie, la plupart des essais cliniques doivent intégrer la qualité de vie relative à la santé (QdV) comme critère de jugement afin d'investiguer le bénéfice clinique pour le patient. La QdV est un concept dynamique dépendant de l'adaptation du patient et reflété par un effet « Response shift ». Cet effet résulte d'un changement dans les références internes du patient (« recalibration »), dans les valeurs (« reprioritization ») et dans la conceptualisation de la QdV (« reconceptualization »). L'analyse longitudinale de la QdV doit tenir compte de l'occurrence éventuelle d'un tel effet. En revanche, il n'existe pas de standard d'analyse statistique pour caractériser cet effet. Deux méthodes complémentaires sont investiguées pour caractériser l'occurrence de la « Response shift ».

Matériels et méthodes

Ce travail s'appuie sur les données d'une cohorte prospective multicentrique incluant toutes les patientes atteintes d'un cancer du sein primitif (ou d'une suspicion). La QdV a été évaluée par les questionnaires EORTC QLQ-C30 et son module QLQ-BR23 spécifique du cancer du sein à l'inclusion, après la chirurgie, à trois mois et à six mois. La recalibration a été explorée selon le design « then-test/post-test »: les évaluations rétrospectives faites post-chirurgie et à trois mois référencent la QdV à l'inclusion; la mesure rétrospective faite à six mois référence la QdV à trois mois. L'ordre de remplissage des questionnaires prospectifs et rétrospectifs a été randomisé. La composante « recalibration » a été explorée par des Analyses Factorielles des Correspondances Multiples (AFCM) et le modèle dit Linear Logistic Model with Relaxed Assumptions (LLRA) de type IRT. Le modèle LLRA traduit la « recalibration » par un changement au niveau du paramètre de facilité de réponse aux items. Les composantes « reprioritization » et « reconceptualization » ont été explorées par des Analyses en Composantes Principales (ACP).

Résultats

Entre février 2006 et février 2008, 381 patientes ont été incluses dont 89% avaient un cancer du sein confirmé. Les ACP révèlent une « reprioritization » secondaire des dimensions du QLQ-C30. Fatigue et douleur restent les symptômes prioritaires. Les symptômes secondaires sont l'insomnie à l'inclusion, la diarrhée après la chirurgie, les nausées et vomissements à trois mois et à six mois. Une « reconceptualization » est illustrée par un lien de plus en plus marqué entre les échelles fonctionnelles. Les principaux profils de « recalibration » reflétés par les AFCM sont d'une modalité de réponse à une modalité de réponse adjacente. Les patientes rapportant peu de symptômes lors de la mesure « pré-test » ont tendance à omettre leur présence lors de la mesure rétrospective « then-test », et inversement. L'application du modèle LLRA indique une « recalibration » à la baisse ou à la hausse de chaque dimension. Concernant la réévaluation de la QdV à trois mois, les symptômes au niveau du bras et du sein sont sous-estimés avec comme paramètre de tendance -1.05 et -0.59 respectivement ($p<0.0001$). Les patientes s'orientent donc vers des modalités plus basses que lors de la mesure prospective.

Conclusion

Les modèles IRT ont surtout été utilisés pour valider des questionnaires de QdV. Ce travail montre leur intérêt dans la caractérisation de l'occurrence de la « Response shift ». Des analyses complémentaires doivent être menées afin d'investiguer et de valider leurs capacités à caractériser toutes les composantes de la « Response shift ».

<u>Article:</u> Item Response Theory and Factor Analysis as a mean to characterize occurrence of Response Shift in a longitudinal quality of life study in breast cancer patients

Article accepté dans *Health and Quality of Life Outcomes*

Anota et al. Health and Quality of Life Outcomes 2014, 12:32
http://www.hqlo.com/content/12/1/32

RESEARCH Open Access

Item response theory and factor analysis as a mean to characterize occurrence of response shift in a longitudinal quality of life study in breast cancer patients

Amélie Anota[1,2,3*], Caroline Bascoul-Mollevi[4], Thierry Conroy[1,5], Francis Guillemin[1,6], Michel Velten[1,7], Damien Jolly[1,8], Mariette Mercier[1,3], Sylvain Causeret[9], Jean Cuisenier[9], Olivier Graesslin[10], Zeinab Hamidou[1,11] and Franck Bonnetain[1,2,3]

Abstract

Background: The occurrence of response shift (RS) in longitudinal health-related quality of life (HRQoL) studies, reflecting patient adaptation to disease, has already been demonstrated. Several methods have been developed to detect the three different types of response shift (RS), i.e. recalibration RS, 2) reprioritization RS, and 3) reconceptualization RS. We investigated two complementary methods that characterize the occurrence of RS: factor analysis, comprising Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA), and a method of Item Response Theory (IRT).

Methods: Breast cancer patients (n = 381) completed the EORTC QLQ-C30 and EORTC QLQ-BR23 questionnaires at baseline, immediately following surgery, and three and six months after surgery, according to the "then-test/post-test" design. Recalibration was explored using MCA and a model of IRT, called the Linear Logistic Model with Relaxed Assumptions (LLRA) using the then-test method. Principal Component Analysis (PCA) was used to explore reconceptualization and reprioritization.

Results: MCA highlighted the main profiles of recalibration: patients with high HRQoL level report a slightly worse HRQoL level retrospectively and vice versa. The LLRA model indicated a downward or upward recalibration for each dimension. At six months, the recalibration effect was statistically significant for 11/22 dimensions of the QLQ-C30 and BR23 according to the LLRA model ($p \leq 0.001$). Regarding the QLQ-C30, PCA indicated a reprioritization of symptom scales and reconceptualization via an increased correlation between functional scales.

Conclusions: Our findings demonstrate the usefulness of these analyses in characterizing the occurrence of RS. MCA and IRT model had convergent results with then-test method to characterize recalibration component of RS. PCA is an indirect method in investigating the reprioritization and reconceptualization components of RS.

Keywords: Health-related quality of life, Longitudinal analysis, Response shift, Factor analysis, Item Response Theory

* Correspondence: aanota@chu-besancon.fr
[1]Quality of Life in Oncology Platform, Besançon, France
[2]Methodological and Quality of Life Unit in Oncology, University Hospital of Besançon, Besançon, France
Full list of author information is available at the end of the article

Background

Health-related quality of life (HRQoL) is a subjective clinical endpoint that has been increasingly important in health outcomes research and particularly in cancer clinical trials over the past two decades [1] as well as in breast cancer [2]. Although overall survival is still considered as the primary objective and the primary endpoint in many studies, most clinical trials now integrate HRQoL as an endpoint in order to investigate the clinical benefit for the patient.

One major objective of measuring HRQoL over time is determining the extent to which treatment toxicities or disease progression can affect patients' HRQoL level. However, self-assessment of HRQoL is subjective, *i.e.* it is dependent on the patient's internal standards and definition of HRQoL [3-5]. As patients can adapt to disease and treatment toxicities, their health and HRQoL expectations can also change over time. These changes could result in a response shift (RS) effect [6-8].

RS can be defined as "a change in the meaning of one's self-evaluation of a target construct as a result of: (a) a change in the respondent's internal standards of measurement (i.e. scale recalibration); (b) a change in the respondent's values (i.e. the importance of component domains constituting the target construct, [reprioritization]) or (c) a redefinition of the target construct (i.e. reconceptualization)" [9].

Different methods have been proposed to assess RS [10-13]. The most widely used is the "Then-test" method, which assesses patients' pre-test HRQoL levels retrospectively. The test involves asking patients post-treatment to provide their current levels (post-test) but also their pre-test levels in retrospect (then-test). This method is based on the assumption that patients rate their HRQoL post-test and pre-test levels with the same criteria, since the assessments occur at the same time point. The recalibration component of RS should thus be taken into account when comparing post-test and then-test scores. Comparing the mean of the pre-test and then-test scores explores recalibration component of RS [12].

Statistical methods have also been investigated to detect RS. First, factor analyses have been explored to detect RS [14,15]. An alternative to investigate RS with factor analysis is the use of structural equation modeling (SEM) [11,16,17]. These models can evaluate all types of RS if they are experienced by a substantial part of individual in the population analyzed [16]. These models are based on means and covariance structures and rely on observed scores. At this time and to our knowledge, these models have never been applied to an European Organization for Research and Treatment of Cancer (EORTC) HRQoL questionnaire in order to highlight RS effect. Principal Component Analysis (PCA) is a special case of SEM. Item Response Theory (IRT) could also be considered to explore RS effect but up to now these models remain less applied and mainly through differential item functioning [18,19]. Contrary to SEM, IRT models are not based on the observed scores but directly on items answers. In fact, in SEM, the raw score is assumed to be a good representation of the latent trait (i.e. HRQoL), while in IRT the items responses play a key role and the relation between the items responses and the latent trait are not linear in IRT.

While occurrence of RS in HRQoL studies has been demonstrated [13], approaches that could reinforce the proof that each component of RS occurs should be investigate to complement the results from other methods. All methods that highlight RS have their strengths and weaknesses then similar trend obtained from different tools should increase accuracy of the results characterizing RS occurrence and the confidence of the results. Moreover, the studies to detect the occurrence of RS are generally performed with two measurement time points while in oncology clinical trials more than two assessments is usually planned. Then we need also tools for the longitudinal analysis of the potential occurrence of RS.

The intent of this study was thus to investigate statistical methods to characterize the occurrence of RS for HRQoL in breast cancer (BC) patients.

The primary objective was to assess if Multiple Correspondence Analysis (MCA), which is a factor analysis and a model of IRT named the Linear Logistic model with Relaxed Assumptions (LLRA), had convergent results with then-test method to characterize recalibration component of RS.

The secondary objective was to assess if Principal Component Analysis (PCA), which is another factor analysis model, could be a valuable tool to longitudinally identify the reconceptualization and reprioritization components of RS independently of the occurrence of the recalibration component of RS.

Methods

Patients and eligibility criteria

A prospective, multicenter, randomized cohort study was performed in the cancer care centers at Dijon, Nancy, and the university hospitals of Strasbourg and Reims (cities of France). It is a collaboration between different teams with complementary skills and an interest in the topic and all these teams are involved in quality of life research. All women initially hospitalized between February 2006 and February 2008 for diagnosis or treatment of primary or suspected BC were eligible for inclusion. We anticipated that patients with no confirmed BC could constitute a control group. Nevertheless, due to the low effective (less than 10% of patients

included) they could not constitute a larger control group then they were excluded from the analyses. Women with cancer other than BC, already undergoing BC treatment, or with a previous history of cancer were excluded. Written informed consent was obtained from every participant and the protocol was approved by Dijon University Hospital Ethics committee [20].

Health-related quality of life assessment

HRQoL was evaluated using the EORTC QLQ-C30 and EORTC QLQ-BR23 BC specific tool at four time points: at baseline (initial examination or initial hospitalization), at discharge following initial hospitalization, at three months (M3) and six months (M6) [21,22]. The QLQ-C30 and its BC module BR23 are validated tools in assessing HRQoL in cancer, specifically in BC [21,22]. The QLQ-C30 comprises 30 items and measures five functional scales (physical, role, emotional, cognitive and social functioning), global health status (GHS), financial difficulties and eight symptom scales (fatigue, nausea and vomiting, pain, dyspnea, insomnia, appetite loss, constipation, diarrhea) [22]. The BR23 module comprises 23 items that generate four functional scales (body image, sexual functioning, sexual enjoyment, future perspective) and four symptom scales (systemic therapy side effects (STSE), breast symptoms, arm symptoms, upset by hair loss) [22].

Response categories vary from 1 to 4 on a Likert scale for the QLQ-C30 and BR23 questionnaires, with 1 corresponding to the best state for functional scales or no symptoms, and 4 corresponding to the worst state for functional scales or the highest symptomatic level. For sexual dimensions, the response categories are reversed. Scores are generated according to the EORTC Scoring Manual [23]. These scores vary from 0 (worst) to 100 (best) for the functional dimensions and GHS, and from 0 (best) to 100 (worst) for the symptom dimensions.

A five-point difference in EORTC HRQoL scores is considered as the minimal clinically important difference (MCID) [24].

Then test assessment

In this study, the retrospective pre-test/post-test design was used to detect recalibration [13]. At each follow-up time point, one prospective and one retrospective measurement were performed. The retrospective assessments at the end of initial hospitalization and at M3 refer to baseline HRQoL. At M6, the retrospective measurement refers to HRQoL at M3. The order of the then-test and post-test of HRQoL questionnaires was randomized with a 1:1 allocation and stratification by center to assess the impact of the order on RS occurrence and estimate. In arm A, the order of the questionnaires was post-test/then-test. In arm B, the order was then-test/post-test. Authorization was obtained from the EORTC HRQoL unit to adapt the HRQoL questionnaires (EORTC QLQ-C30 and module BR23) to the then-test assessment. The impact of the retrospective or prospective administration of the questionnaire on RS occurrence has already been analyzed in a previous paper showing no order effect and is not treated in the present paper [20].

Treatments as well as clinical and sociodemographic variables were recorded at inclusion.

Statistical analyses

Studied population and missing data

Variables collected at baseline were described with median and range for continuous variables and percentage for qualitative variables, with percentage of missing data. No imputation was performed on missing items. Scores were calculated if at least half of the items were answered according to the recommendations of the EORTC scoring manual [23]. No imputation was performed on missing scores.

MCA and LLRA were both performed on patients with all items of the studied dimension (each dimension of the QLQ-C30 and the QLQ-BR23) filled out at the then-test and the pre-test measurement time points and with a MCID between then-test and pre-test of at least 5 points for the given dimension. This selection was done in order to retain a clinically meaningful difference of the recalibration occurrence.

PCA were performed on patients with all scores available at the four prospective measurement times for one questionnaire (QLQ-C30 or BR23).

For each analysis, patients retained were compared to those excluded according to baseline characteristics in order to check the random missing data profile and then a possible selection bias.

Recalibration

For each score, the mean difference (MD) between each then-test and the corresponding pre-test was calculated and described as mean (SD). The existence of a significant recalibration was tested with a Wilcoxon matched pairs test. The effect size was calculated in order to assess the magnitude of RS effect and was defined as the mean change score between the then-test and the corresponding pre-test dividing by the standard deviation of patients at the prospective measurement time.

The primary objective was to assess if MCA and the LLRA model of IRT had convergent results with the then-test method to characterize the recalibration component of RS.

Firstly, recalibration was thus explored by MCA [25]. MCA is a factor analysis dedicated to qualitative variables and can identify links between categories of polytomous variables. This method is thus well adapted to the items constructed on a Likert scale. This analysis was

applied to items of each dimension according to the Then-test method, i.e. with pre-test and then-test measures of the same HRQoL. Only recalibration was explored with this method since only one dimension was included. Therefore, recalibration was confirmed by a correlation between two different response categories of the same item measured at pre-test and at then-test measurement time [26]. The study was limited to the first two axes.

LLRA, a IRT model for measuring change, was then applied to explore recalibration [27-30].

IRT and Classical Test Theory differ in terms of score calculation. Classical Test Theory is mainly based on observed scores while in IRT, item responses play the key role: IRT models the item responses to the latent trait by a probabilistic model. The raw score is thus not considered as a good representation of the latent trait but the response to each item is considered directly. The relationship between the observed score and the latent trait is no longer linear. They are generally linked and modeled by a logistic function. The IRT models introduce the concepts of item easiness parameters and person parameters.

The person parameter corresponds to the level of the patient on the latent trait (e.g. the level of HRQoL). The item parameter is the location of the item on the latent trait and corresponds to a level of difficulty or easiness in this model.

The LLRA requires neither items' unidimensionality nor distributional assumptions about the population of subjects [31]. In addition, the LLRA can fit with polytomous responses and was developed in order to measure the change occurring between several measurement time points [32]. To give up the unidimensionality of the items, items have to be measured at two measurement time points or more [32,33].

The main idea of the LLRA model is not to consider longitudinal change as a change in person parameters, but rather as a change in item parameters. In this way, person parameters are fixed over time and only item parameters vary. Since person parameters are nuisance parameters, we can estimate the item parameter trend instead of the person parameter trend by conditional maximum likelihood [34]. Indeed, fewer parameters have to be estimated and they did not depend on the sample considered.

One item I with parameter βi evaluated twice on an individual can be seen as a pair of virtual items I*1 and I*2 with two item parameters β^*i1 and β^*i2 respectively. For the pre-test, $\beta^*i1 = \beta i$ while for the then-test $\beta^*i2 = \beta i + \tau$ where τ is the upward or downward trend effect of item easiness parameter. This parameter is targeted by LLRA [35]. In cases of polytomous items, for each item with $(m + 1)$ response categories there are m category parameters. The trend parameter τ is the same for each category parameter. The design matrix was constructed such that there is one trend parameter for each item. If possible, the trend was generalized for all items of a dimension. The general form of LLRA, a longitudinal IRT model adapted to polytomous items, is based on the partial credit approach [35].

A positive (or negative) trend τ for one item implies that the item easiness parameter increases (or decreases) at the time of the then-test measurement compared to the pre-test measurement. Patients choose higher (or lower) response categories in the retrospective then-test measure than in the prospective one for this item. In this way, recalibration would be indicated by a significant positive or negative trend for one dimension.

Convergent results between MCA and IRT would correspond to:

- a significant positive trend parameter for IRT and some upward recalibration profiles highlighted by MCA (i.e. patients choose upper response categories at the then-test assessment as compared to the prospective measurement time) more than some downward recalibration profiles (patients choose lower response categories at the then-test assessment as compared to the prospective measurement time).
- a significant negative trend parameter for IRT and some downward recalibration profiles highlighted by MCA (patients choose lower response categories at the then-test assessment) rather than some upward recalibration profiles (patients choose upper response categories at the then-test assessment).
- an insignificant trend parameter for IRT and well-balanced recalibration profiles highlighted by MCA (as many patients choose higher than lower response categories at the retrospective measurement time as compared to the prospective measure).

GHS was excluded from MCA and LLRA because of the high number of response categories. There are seven response categories for both items measuring GHS. To apply a longitudinal model of IRT, all seven categories have to be represented at each measurement time point, which was not the case in the present study. GHS was excluded from MCA in order to be consistent with IRT.

Reprioritization and reconceptualization

The secondary objective was to assess if PCA could be a valuable tool to longitudinally identify the reconceptualization and reprioritization components of RS independently of the occurrence of recalibration component of RS.

PCA was performed on patients with all scores available at all prospective measurement times and for one questionnaire (QLQ-C30 or BR23) on the scores generated for all dimensions of each prospective questionnaire [12,14,15,36]. PCA was performed only for one questionnaire in order to

have clear and understandable graphs. Reprioritization was indicated by a change in scores generating the first two principal components: scales generating the first principal component are a priority to patients while those generating the second principal component are secondary. Changes occurring at the first principal component are considered as major and those occurring at the second principal component as minor. In this way, changes were qualified in the first axis of "major reprioritization" and in the second axis of "secondary reprioritization". The study was limited to the first two principal components, according to the Scree test [37]. Reconceptualization was reflected by a change in the structure of the graph of correlations between scores and principal components, as well as in the connection or opposition of some scores. Concerning the module BR23, sexual enjoyment and hair loss were excluded from the analysis given the number of missing values.

All analyses were performed with R software [38] using FactoMineR library for factor analyses [39] and eRm library for LLRA [34,35,40].

The statistical significance level was reduced to $p = 0.002$ for all analyses in order to prevent false positive results due to the number of multiple comparisons performed (alpha risk 0.05 divided by the number of dimensions analyzed).

Results

Patients

Between February 2006 and February 2008, 381 patients were included in the four participating centers. Mean age was 58.4 (standard deviation = 11) years. Three hundred and forty (89%) patients had confirmed BC. Complete clinical and pathologic features of the population are given in Table 1.

Table 1 Baseline patient characteristics

	N	%
Hospital		
Dijon	271	71.1
Nancy	74	19.4
Reims	18	4.7
Strasbourg	18	4.7
Inclusion criteria		
Confirmed primary breast cancer	242	63.5
Suspicion of primary breast cancer	138	36.2
Unknown	1	0.3
Cancer		
Confirmed	340	89.2
Not confirmed	38	10.0
Unknown	3	0.8
Lymph node dissection(LND)		
Axillary LND	138	36.2
Sentinel lymph node biopsy	131	34.4
ALND + SLNB	32	8.4
No LND	75	19.7
Unknown	5	1.3
Surgery type		
Mastectmoy	124	32.6
No mastectomy	241	63.3
Unknown	16	4.2
Chemotherapy		
Yes	155	40.7
No	218	57.2
Unknown	8	2.1
Radiotherapy		
Yes	254	66.7
No	119	31.2
Unknown	8	2.1
Hormone therapy		
Yes	170	44.6
No	203	53.3
Unknown	8	2.1
Questionnaires order		
Arm 1: then-test/post-test	192	50.4
Arm 2: post-test/then-test	189	49.6

HRQoL questionnaires completion and missing data

Table 2 describes the number of completed QLQ-C30 and BR23 questionnaires at each measurement time.

317 (93%) patients had at least one HRQoL score at baseline, 311 (91%) on discharge following initial hospitalization (i.e. after surgery), 304 (89%) at M3 and 290 (85%) at M6.

Median time for HRQoL assessments between baseline and the discharge following initial hospitalization was 6 days, range [1.5; 81.5].

Patients retained for MCA and LLRA with a 5-point MCID were similar to those excluded according to baseline characteristics for each analysis (data not shown). Patients retained for PCA with all the four prospective measurement times were similar to those excluded except that they seem to be older (data not shown).

Recalibration

After surgery (Table 3), the recalibration effect was statistically and clinically significant for emotional functioning (MD = 5.36) and future perspectives (MD = 7.41) dimensions with a moderate effect size (0.21 and 0.24 respectively). At M3, the recalibration effect was statistically and clinically significant for role (MD = –6.50), emotional (MD = 6.97) and social functioning (MD = –5.01), insomnia (MD = –6.93), body image (MD = –8.16) and future perspectives (MD = 6.95) dimensions.

Table 2 Description of the EORTC QLQ-C30 and BR23 questionnaires received at each measurement time

	QLQ-C30		QLQ-BR23	
	Then-test	Post-test	Then-test	Post-test
Baseline		359 (94.2%)		357 (93.7%)
After surgery	347 (91.1%)	347 (91.1%)	347 (91.1%)	346 (90.8%)
3 months	339 (90.0%)	342 (89.8%)	355 (87.9%)	340 (89.2%)
6 months	314 (82.4%)	322 (84.5%)	313 (82.1%)	322 (84.5%)

At M6, the recalibration effect was statistically and clinically significant for physical (MD = 5.10), role (MD = 8.55) and social functioning (MD = 6.02) and for fatigue (MD = −11.03), pain (MD = −6.02), insomnia (MD = −5.64), body image (MD = 7.78) and breast symptoms (MD = −7.28).

Recalibration and MCA

All results obtained on the QLQ-C30 and QLQ-BR23 are summarized in Table 4 and in Table 5, respectively. Qi_k (resp. Ri_k) refers to the k-th response category of the i-th item of a prospective (resp. retrospective) questionnaire on the graph.

Figure 1 presents the graph obtained for baseline and the then-test performed after surgery for role functioning. 272 patients answered the items 6 (Were you limited in doing either your work or other daily activities?) and 7 (Were you limited in pursuing your hobbies or other leisure time activities?) measuring the role functioning scale at baseline and at the retrospective measurement after surgery referring to the baseline HRQoL. Response categories are coded "1/2/3/4" respectively for "Not at all/A little/Quite a bit/Very much". Among these patients, 84 (31%) had a MCID of at least 5 points between the two measures. Figure 1 highlights two main patterns of recalibration: patients who had reported an excellent role functioning at baseline (i.e. had chosen response category 1 for both items 6 and 7 at baseline) and who had declared a slightly worse role functioning level when they reevaluated this dimension retrospectively after surgery (i.e. chose response category 2 for both items 6 and 7 at the retrospective measurement time), and vice versa (i.e. had chosen response category 2 for both items 6 and 7 at baseline and had chosen response category 2 for both items 6 and 7 at the retrospective measurement time). The first profile is suggested by an association between Q6_1, Q7_1, R6_2 and R7_2. The reverse profile corresponds to the association between Q6_2, Q7_2, R6_1 and R7_1. Recalibration profiles are less explicit for patients who had reported a low role functioning at baseline (i.e. had chosen response category 3 or 4 for both items 6 and 7 at baseline). Indeed, these patients are fewer and they did not follow a unique recalibration profile. Patients who had reported a relatively low role functioning at baseline (i.e. had chosen response category 3 for both items 6 and 7 at baseline) either tended to revise their opinion upwards or downwards by choosing either response category 4 or response category 2 for both items 6 and 7 at the retrospective assessment after surgery.

Recalibration and IRT using LLRA

Positive trend parameters ($\tau = +$) indicated that at the pre-test measurement, patients had overestimated their functional level or had underestimated their symptomatic or sexual level.

Based on the first retrospective reassessment of their baseline HRQoL (Table 6), patients had significantly underestimated their emotional ($\tau = -0.62$, $p < 0.001$) and cognitive functioning ($\tau = -1.15$, $p < 0.001$) and their level of arm symptoms ($\tau = 0.64$, $p < 0.001$). Patients also had overestimated the presence of insomnia ($\tau = -0.49$, $p = 0.001$) and diarrhea ($\tau = -1.34$, $p < 0.001$).

Based on the second retrospective reassessment of their baseline HRQoL, patients had significantly overestimated their role ($\tau = 0.71$, $p < 0.001$) and social functioning ($\tau = 0.66$, $p < 0.001$), their body image ($\tau = 0.83$, $p < 0.001$) and insomnia level ($\tau = -0.49$, $p < 0.001$) and had underestimated their level of pain ($\tau = 0.37$, $p = 0.001$), and arm symptoms ($\tau = 0.86$, $p < 0.001$).

Regarding HRQoL at M3, patients had significantly underestimated their physical ($\tau = -0.84$, $p < 0.001$), role ($\tau = -0.60$, $p < 0.001$), cognitive ($\tau = -0.53$, $p < 0.001$) and social functioning ($\tau = -0.51$, $p < 0.001$) as well as their body image ($\tau = -0.66$, $p < 0.001$). Patients also had overestimated their emotional functioning ($\tau = 0.22$) and their levels of fatigue ($\tau = -0.54$), pain ($\tau = -0.54$), arm ($\tau = -0.37$) and breast ($\tau = -0.76$) symptoms ($p \leq 0.001$).

To summarize, after surgery, the recalibration effect was statistically significant for 6/22 dimensions of the QLQ-C30 and BR23 according to the IRT model while for the then-test method it was only clinically significant for 2 of these dimensions (emotional functioning and future perspective). At M3, the recalibration effect was statistically significant for 7/22 dimensions of the QLQ-C30 and BR23 according to the IRT model and to the then-test method except for the arm symptoms (MD = 2.43, $p = 0.011$). At M6, the recalibration effect was statistically significant for 11/22 dimensions of the QLQ-C30 and BR23 according to the IRT model. The same results were observed for the then-test method except for the emotional ($p = 0.054$) and cognitive functioning ($p = 0.004$) and the arm symptoms ($p = 0.010$). A significant and clinically recalibration was also observed according to the classical then-test method for insomnia (MD = −5.64, $p = 0.002$) and not according to the IRT model ($p = 005$).

Table 3 Recalibration component of response shift effect assessed with the then-test method at each measurement time

	Baseline HRQoL		Then-test 1 minus pre-test				Then-test 2 minus pre-test				HRQoL at three months		Then-test 3 minus pre-test			
	N	Mean (SD)	N	Mean (SD)	*P*	Effect size	N	Mean (SD)	*P*	Effect size	N	Mean (SD)	N	Mean (SD)	*P*	Effect size
QLQ-C30																
Global Health Status	310	68.66 (20.52)	280	−0.80 (16.67)	0.600	−0.04	275	**−4.21 (18.45)**	**<0.001**	**−0.21**	300	60.02 (20.19)	266	0.91 (21.30)	0.48	0.05
Physical functioning	313	90.01 (15.50)	283	−0.50 (10.78)	0.987	−0.03	280	−1.59 (13.26)	0.053	−0.11	301	81.31 (16.76)	273	**5.10 (14.52)**	**<0.001**	**0.31**
Role functioning	309	89.28 (20.38)	281	−1.90 (18.70)	0.182	−0.09	282	**−6.50 (23.72)**	**<0.001**	**−0.32**	300	74.06 (30.16)	269	**8.55 (28.99)**	**<0.001**	**0.30**
Emotional functioning	308	64.86 (26.20)	279	**5.36 (18.64)**	**<0.001**	**0.21**	280	**6.97 (21.48)**	**<0.001**	**0.27**	300	72.80 (22.54)	270	−2.85 (24.55)	0.054	−0.11
Cognitive functioning	313	83.23 (20.76)	280	2.80 (14.97)	0.027	0.13	281	2.37 (18.27)	0.041	0.11	301	82.84 (21.11)	239	4.03 (20.80)	0.004	0.17
Social functioning	307	90.34 (18.88)	264	−0.51 (16.18)	0.644	−0.03	276	**−5.01 (20.70)**	**<0.001**	**−0.27**	298	81.37 (25.42)	266	**6.02 (25.77)**	**<0.001**	**0.22**
Fatigue	310	22.89 (22.92)	278	−1.48 (18.22)	0.039	−0.06	279	1.75 (20.92)	0.228	0.08	300	32.82 (24.08)	270	**−11.03 (25.36)**	**<0.001**	**−0.43**
Nausea and vomiting	312	3.53 (11.18)	270	−0.77 (8.34)	0.130	−0.07	282	1.77 (15.11)	0.092	0.16	299	3.44 (10.90)	269	−3.22 (19.95)	0.010	−0.18
Pain	316	12.45 (20.87)	285	0.53 (19.04)	0.897	0.02	283	3.24 (23.03)	0.032	0.15	304	25.08 (24.82)	274	**−6.02 (23.98)**	**<0.001**	**−0.23**
Dyspnea	310	11.72 (31.87)	280	−2.02 (15.19)	0.036	−0.09	279	−1.08 (15.58)	0.333	−0.05	301	12.86 (20.89)	269	−3.59 (24.08)	0.009	−0.15
Insomnia	307	38.11 (31.87)	277	−5.30 (27.14)	0.003	−0.17	274	**−6.93 (30.94)**	**<0.001**	**−0.22**	299	36.63 (30.27)	266	**−5.64 (32.59)**	**0.002**	**−0.18**
Appetite loss	312	11.75 (22.46)	280	−3.45 520.35	0.005	−0.15	280	−1.19 (23.75)	0.323	−0.05	299	10.20 (20.13)	267	−4.62 (25.02)	0.004	−0.18
Constipation	310	12.47 (22.80)	276	−1.09 (21.15)	0.520	−0.05	277	1.56 (24.93)	0.284	0.07	298	21.38 (30.87)	264	−4.29 (26.61)	0.006	−0.16
Diarrhea	309	8.63 (16.48)	278	−2.88 (12.66)	<0.001	−0.17	277	−2.89 (17.25)	0.010	−0.17	296	4.76 (12.30)	263	−0.63 (22.07)	0.483	−0.04
Financial difficulties	300	4.56 (14.60)	264	0.38 (12.83)	0.741	0.03	269	0.99 (16.51)	0.453	0.07	299	5.86 (15.93)	264	−2.15 (18.14)	0.048	−0.10
QLQ-BR23																
Body image	295	90.04 (17.32)	262	−0.76 (11.95)	0.505	−0.05	257	**−8.16 (16.96)**	**<0.001**	**0.48**	302	70.76 (30.77)	269	**7.78 (24.82)**	**<0.001**	**0.25**
Sexual functioning	274	76.46 (24.01)	232	−0.50 (13.82)	0.207	−0.02	222	−1.21 (15.92)	0.206	−0.05	267	79.65 (22.06)	224	−4.69 (18.52)	0.002	−0.21
Sexual enjoyment	126	43.92 (28.79)	99	2.36 (11.91)	0.124	0.09	100	3.33 (22.97)	0.284	0.12	138	52.17 (29.31)	108	−1.54 (23.41)	0.548	−0.06
Future perspective	295	47.46 (30.86)	261	**7.41 (30.60)**	**<0.001**	**0.24**	259	**6.95 (32.47)**	**<0.001**	**0.23**	301	54.49 (32.76)	269	−0.12 (33.02)	0.968	−0.01
STSE	308	13.29 (15.30)	280	−1.71 (9.89)	0.008	−0.11	271	0.73 (13.84)	0.563	0.05	301	25.13 (20.20)	271	**−4.76 (19.52)**	**<0.001**	**−0.24**
Breast symptoms	273	11.25 (14.92)	243	−0.73 (14.39)	0.152	−0.05	239	2.15 (19.76)	0.379	0.14	302	24.73 (22.97)	273	**−7.28 (20.70)**	**<0.001**	**−0.31**
Arm symptoms	297	8.06 (14.42)	268	1.58 (18.49)	0.572	0.11	261	2.43 (16.53)	0.011	0.17	302	16.39 (18.54)	273	−2.71 (17.70)	0.010	−0.15
Hair loss	55	32.73 (36.57)	34	0.98 (17.38)	0.749	0.03	31	1.08 (25.07)	0.506	0.03	131	53.69 (39.78)	54	−8.03 (40.92)	0.173	−0.21

P: Wilcoxon matched test *P* value.
SD: standard deviation.
Results in bold correspond to clinically and statistically significant results.

Table 4 Main recalibration profiles highlighted by a multiple correspondence analysis performed on the EORTC QLQ-C30

Dimension (items)	Time points	N	Percentage of recalibration (number of patients)	Recalibration category 1 to category 2	Recalibration category 2 to category 1	Recalibration category 3 to category 4	Recalibration category 4 to category 3	Categories 3 and 4 dispersed
Physical functioning	T1 - T2_R[a]	100	37% (272)	Q1-Q5	Q1-Q5			Q1-Q5
(Q1, Q5)	T1 - T3_R[b]	139	51% (274)	Q1-Q5	Q1-Q5			Q1-Q5
	T3 - T4_R[c]	201	76% (266)	Q1-Q5	Q1-Q5			Q1-Q5
Role functioning	T1 - T2_R	84	31% (272)	Q6, Q7	Q6, Q7			Q6, Q7
(Q6-Q7)	T1 - T3_R	118	43% (274)	Q6, Q7	Q6, Q7	Q6, Q7	Q6, Q7	
	T3 - T4_R	164	63% (261)	Q6, Q7	Q6, Q7			Q6, Q7
Emotional functioning	T1 - T2_R	180	68% (263)	Q21-Q24	Q21-Q24			
(Q21-Q24)	T1 - T3_R	208	79% (263)	Q21-Q24	Q21-Q24			
	T3 - T4_R	196	77% (255)	Q21-Q24	Q21-Q24			Q21-Q24
Cognitive functioning	T1 - T2_R	103	39% (266)	Q20, Q25				
(Q20, Q25)	T1 - T3_R	129	48% (268)	Q20, Q25	Q20, Q25	Q20, Q25	Q20, Q25	Q20, Q25
	T3 - T4_R	130	51% (256)	Q20, Q25	Q20, Q25	Q20, Q25	Q20, Q25	
Social functioning	T1 - T2_R	75	29% (260)	Q26, Q27	Q26, Q27			Q26, Q27
(Q26, Q27)	T1 - T3_R	101	38% (268)	Q26, Q27	Q26, Q27	Q26, Q27	Q26, Q27	
	T3 - T4_R	142	56% (255)	Q26, Q27	Q26, Q27	Q26, Q27	Q26, Q27	Q26, Q27
Fatigue	T1 - T2_R	141	54% (259)	Q10, Q12, Q18	Q10, Q12, Q18			Q10, Q12, Q18
(Q10, Q12, Q18)	T1 - T3_R	160	61% (261)	Q10, Q12, Q18	Q10, Q12, Q18	Q12	Q10, Q12, Q18	Q10, Q12, Q18
	T3 - T4_R	183	73% (251)	Q10, Q12, Q18	Q10, Q12, Q18		Q10, Q12, Q18	Q10, Q12, Q18
Nausea and vomiting	T1 - T2_R	37	13% (280)	Q14	Q14, Q15			Q14, Q15
(Q14, Q15)	T1 - T3_R	65	24% (275)	Q14	Q14		Q14	
	T3 - T4_R	98	37% (265)	Q14	Q14			
Pain	T1 - T2_R	96	36% (267)	Q9	Q9			Q9, Q19
(Q9, Q19)	T1 - T3_R	124	46% (268)	Q9, Q19	Q9, Q19			Q9, Q19
	T3 - T4_R	148	58% (253)	Q9, Q19	Q9, Q19		Q9, Q19	
Insomnia	T1 - T2_R	115	42% (277)	Q11	Q11		Q11	
(Q8)	T1 - T3_R	135	49% (274)	Q11	Q11	Q11	Q11	
	T3 - T4_R	147	55% (266)	Q11	Q11	Q11	Q11	
Dyspnea	T1 - T2_R	50	18% (280)	Q8	Q8		Q8	
(Q11)	T1 - T3_R	55	20% (279)	Q8	Q8	Q8	Q8	
	T3 - T4_R	94	35% (269)	Q8	Q8	Q8	Q8	

Table 4 Main recalibration profiles highlighted by a multiple correspondence analysis performed on the EORTC QLQ-C30 *(Continued)*

Appetite loss	T1 - T2_R	61	22% (280)	Q13	Q13		Q13
(Q13)	T1 - T3_R	85	30% (280)	Q13	Q13		Q13
	T3 - T4_R	90	34% (267)	Q13	Q13		Q13
Constipation	T1 - T2_R	71	26% (276)	Q16	Q16	Q16	Q16
(Q16)	T1 - T3_R	89	32% (277)	Q16	Q16	Q16	Q16
	T3 - T4_R	93	35% (264)	Q16	Q16	Q16	Q16
Diarrhea	T1 - T2_R	39	14% (278)	Q17	Q17		
(Q17)	T1 - T3_R	64	23% (277)	Q17	Q17		
	T3 - T4_R	61	24% (263)	Q17	Q17		Q17
Financial difficulties	T1 - T2_R	22	8% (264)	Q28	Q28		
(Q28)	T1 - T3_R	33	12% (269)	Q28	Q28		Q28
	T3 - T4_R	47	18% (264)	Q28	Q28		Q28

Only patients with a recalibration of 5-point at least between a pre-test and a then-test measure are incorporated in these analyses. Items with observed recalibration are listed.
[a]T1 → T2_R: comparison of baseline HRQoL assessment and retrospective measure performed after surgery.
[b]T1 → T3_R: comparison of baseline HRQoL assessment and retrospective measure performed three months later.
[c]T3 → T4_R: comparison of HRQoL assessment at three months and retrospective measure performed three months later.
As example, 100 patients presented a significant recalibration of physical functioning among the 272 patients with all the five items of the dimension answered at both measurement time. The graph representing the response categories highlight some recalibration profile:
- patients who had chosen response category 1 at the prospective measurement time for all the 5 items and who had chosen response category 2 to the same items at the retrospective measurement time.
- patients who had chosen response category 2 at the prospective measurement time for all the 5 items and who had chosen response category 1 to the same items at the retrospective measurement time.
- no recalibration profile is highlighted for response categories 3 and 4 and few patients had chosen these categories of response (these response categories are dispersed).

Table 5 Main recalibration profiles highlighted by multiple correspondence analysis performed on the QLQ-BR23 for patients with recalibration

		N	Percentage of recalibration (number of patients retained)	Recalibration category 1 to category 2	Recalibration category 2 to category 1	Recalibration category 3 to category 2	Recalibration category 3 to category 4	Recalibration category 4 to category 3	Categories 3 and 4 dispersed
Body image	T1-T2_R[a]	68	29% (236)	Q9-Q12	Q9-Q12			Q9, Q10	Q9-Q12
(Q9-Q12)	T1-T3_R[b]	119	50% (238)	Q9-Q12	Q9-Q12		Q9, Q12		Q9-Q12
	T3-T4_R[c]	153	63% (244)	Q9, Q10	Q9, Q10				Q9-Q12
Sexual functioning	T1-T2_R	55	25% (219)	Q14		Q14			
(Q14, Q15)	T1-T3_R	71	34% (210)	Q14		Q14, Q15			
	T3-T4_R	78	37% (213)	Q14, Q15	Q14, Q15	Q14, Q15			
Sexual enjoyment	T1-T2_R	13	13% (99)	Q16	Q16	Q16	Q16	Q16	
(Q16)	T1-T3_R	39	39% (100)	Q16	Q16	Q16	Q16	Q16	
	T3-T4_R	33	31% (108)	Q16	Q16	Q16	Q16	Q16	
Future perspectives	T1-T2_R	122	47% (261)	Q13	Q13	Q13	Q13	Q13	
(Q13)	T1-T3_R	135	52% (259)	Q13	Q13		Q13	Q13	
	T3-T4_R	144	54% (269)	Q13	Q13		Q13	Q13	
Systemic therapy side effects	T1-T2_R	55	26% (209)						
(Q1-Q4, Q6-Q8)	T1-T3_R	61	32% (190)						
	T3-T4_R	77	39% (200)						
Breast symptoms	T1-T2_R	114	51% (223)	Q20-Q23	Q20-Q23				Q20-Q23
(Q20-Q23)	T1-T3_R	135	61% (223)	Q20-Q23	Q20-Q23	Q20-Q23		Q20-Q23	Q20-Q23
	T3-T4_R	190	74% (258)	Q20-Q23	Q20-Q23	Q20-Q23			Q20-Q23
Arm symptoms	T1-T2_R	97	38% (255)	Q17-Q19	Q17-Q19				Q17-Q19
(Q17-Q19)	T1-T3_R	111	46% (244)	Q17-Q19	Q17-Q19			Q17	Q17-Q19
	T3-T4_R	156	60% (260)	Q17	Q17-Q19				Q17-Q19
Hair loss	T1-T2_R	6	18% (34)	Q5		Q5		Q5	
(Q5)	T1-T3_R	14	45% (31)	Q5		Q5	Q5	Q5	
	T3-T4_R	24	44% (54)	Q5	Q5		Q5	Q5	

Only patients with a recalibration of 5-point at least between a pre-test and a then-test measure are incorporated in these analyses. Items with observed recalibration are listed.
[a]T1 → T2_R: comparison of baseline HRQoL assessment and retrospective measure performed after surgery.
[b]T1 → T3_R: comparison of baseline HRQoL assessment and retrospective measure performed three months later.
[c]T3 → T4_R: comparison of HRQoL assessment at three months and retrospective measure performed three months later.

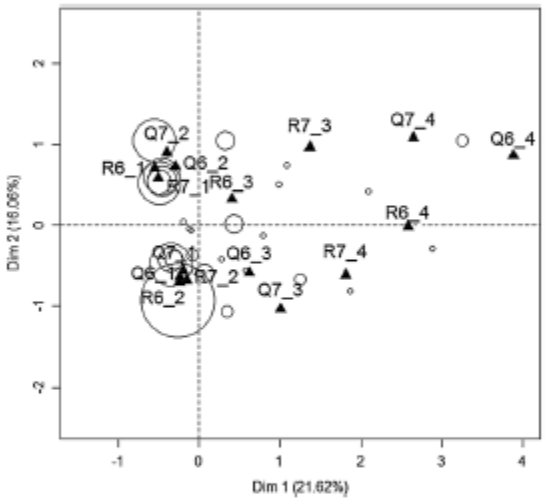


Figure 1 Variable response categories according the first two axes obtained by Multiple Correspondence Analysis between prospective measure of role functioning at baseline and retrospective measure after surgery (N = 84). Questions 6 and 7 of the EORTC QLQ-C30 measure this dimension: Qi_k (or Ri_k) refers to the k-th category of the i-th item of a prospective (or retrospective) questionnaire on the graph. Patients are represented by circles that are proportional to the number of patients with the same coordinates (i.e. who had given the same responses).

Reprioritization and reconceptualization with PCA

PCA was performed on each prospective measure performed at baseline, post-surgery, and at M3 and M6, on patients with all scores available at the prospective measurement time for one questionnaire (QLQ-C30 or QLQ-BR23).

For the QLQ-C30 (respectively QLQ-BR23), 192 (50.4%) patients (respectively 154 (40.4%)) were retained with all scores available at the four prospective measurement times.

Concerning the QLQ-C30 (Figure 2), functional scales became more interrelated and related to the first principal component, reflecting a strong positive correlation between these scales (Table 7). This is observed at each measurement time point. Fatigue and pain remained strongly correlated at each measurement time point, a little less at M3. Diarrhea and financial difficulties were correlated just after surgery (Figure 2B). Nausea and vomiting were correlated to appetite loss at M6 (Figure 2D).

Reprioritization was mainly secondary, as it mostly affected the second principal component: fatigue and pain were still priority symptoms at each prospective measure, since they were still highly correlated with the first principal component at each prospective measure. These symptoms mainly affected physical, social and role functioning as well as GHS. At M6, all functional scales were affected by these symptoms. The second axis highlighted secondary symptoms, namely insomnia at baseline, diarrhea after surgery, nausea and vomiting at M3 and M6 (Table 7).

Concerning the QLQ-BR23, STSE affected the patients' body image since these dimensions remained strongly negatively correlated (Figure 3). These scales were highly correlated with the first principal component (Table 7). Post-surgery, arm symptoms as well as body image and STSE were significantly correlated with the first principal component. Thus, arm symptoms were equally important to body image and STSE regarding patient HRQoL level. At M3 and M6, future perspectives became significantly correlated with the first principal component and thus gained importance, highlighting a major reprioritization. Breast symptoms and sexual functioning were not significantly associated with the two first axes at baseline: they were minor dimensions. After surgery, only sexual functioning was a relevant factor, while at M3 and M6, only breast symptoms were relevant.

Reconceptualization is illustrated by changes in correlations (positive or negative) between variables at each measurement time point. At each measurement time point, body image score was opposed to STSE score. Post-surgery, STSE was associated with arm symptoms. At M3 and M6, a high body image score was associated with high future perspective.

Discussion

The present study demonstrates that response shift effect occurred in patients with primary breast cancer, just after surgery, as well as at 3 and 6 months. The intent of this study was to investigate statistical methods to characterize the occurrence of response shift in breast cancer patients.

Our primary objective was to assess if MCA and IRT model had convergent results with the then-test method to characterize recalibration component of RS.

Both methods explored are convergent to the then-test method. When the then-test method highlighted a clinically significant recalibration, MCA highlighted a general trend to overestimate or underestimate their HRQoL level choosing higher or lower response categories according to the direction of the recalibration effect. IRT model showed a statistically significant general trend (positive or negative) of item easiness parameter with the exception of insomnia at 6 months for which a recalibration is not detected by IRT at the alpha level $p = 0.002$ but borderline ($p = 0.005$). When the mean difference between the then-test and the pre-test is not significant, i.e. no clinically significant recalibration occurs, MCA highlighted as many patients recalibrate upward than downward their HRQoL level and then there is not a general trend to overestimate or underestimate their HRQoL level. The IRT model also highlighted that the trend of item easiness parameter is not significant. However, some discrepancies are observed: for 4/6 dimensions for which a recalibration was detected by IRT and

Table 6 Trend τ of item easiness parameter estimated by linear logistic model with relaxed assumptions for each quality of life dimension

Dimension	Items	T1 → T2_R[a]			T1 → T3_R[b]			T3 → T4_R[c]		
		N	τ	p	N	τ	p	N	τ	p
QLQ-C30										
Physical functioning	1 – 4[d]	100	−0.02	0.858	139	0.27	0.004	201	−0.84	<0.001
Role functioning	6, 7	84	0.36[e]	0.011	118	0.71	<0.001	164	−0.60	<0.001
Emotional functioning	21 - 24	180	−0.62	<0.001	208	−0.65	<0.001	196	0.22	<0.001
Cognitive functioning	20, 25	103	−1.15	<0.001	129	−0.29	0.020	130	−0.53	<0.001
Social functioning	26, 27	75	0.07	0.791	101	0.66	<0.001	142	−0.51	<0.001
Fatigue	10, 12, 18	141	−0.19	0.280	160	0.23	0.019	183	−0.99	<0.001
Nausea and vomiting	14, 15	37	−0.60	0.022	65	0.50	0.029	98	−0.36	0.003
Pain	9, 19	96	0.11	0.410	124	0.37	0.001	148	−0.54	<0.001
Dyspnea	8	50	−0.60	0.025	55	−0.30	0.248	94	−0.42	0.014
Insomnia	11	115	−0.49	0.001	135	−0.49	<0.001	147	−0.36	0.005
Appetite loss	13	61	−0.59	0.004	85	−0.14	0.401	90	−0.51	0.003
Constipation	16	71	−0.16	0.392	89	0.17	0.295	93	−0.41	0.009
Diarrhea	17	39	−1.34	<0.001	64	−0.67	0.005	61	−0.09	0.641
Financial difficulties	28	22	0.16	0.630	33	0.25	0.322	47	−0.45	0.053
QLQ-BR23										
Body image	9 - 12	68	0.10	0.489	119	0.83	<0.001	153	−0.66	<0.001
Sexual functioning	14, 15	55	0.06	0.950	71	0.33	0.078	78	0.77	0.010
Sexual enjoyment	16	13	−1.20	0.04	39	−0.43	0.14	33	0.19	0.49
Future perspective	13	122	−0.56	<0.01	135	−0.46	<0.01	144	0.01	0.95
STSE	1 - 4; 6 - 8	55	−0.40	<0.01	61	0.21	0.013	77	−0.74	<0.001
Breast symptoms	20 - 23	114	−0.11	0.07	135	0.36	0.01	190	−0.76	<0.001
Arm symptoms	17 - 19	97	0.64	<0.001	167	0.86	<0.001	156	−0.37	0.001
Hair loss	Q5	6	0.22	0.738	16	0.12	0.808	24	−0.33	0.149

[a]T1 → T2_R: comparison of baseline quality of life assessment and retrospective measure performed after surgery.
[b]T1 → T3_R: comparison of baseline quality of life assessment and retrospective measure performed three months later.
[c]T3 → T4_R: comparison of quality of life assessment at three months and retrospective measure performed three months later.
[d]Q5 was excluded because of the patients' responses to this item at baseline: all patients chose either category 1 or 2 for this item at baseline, whereas all four categories were represented on retrospective measures.
[e]τ =0.364: patients significantly chose higher response categories at the retrospective measure of baseline HRQoL performed after surgery.

not significant according to the then-test method at the first retrospective assessment, 1/7 at the second one, and 3/11 at the last retrospective assessment. Thus, the IRT model detects more recalibration effect than the classical then-test method.

MCA and IRT models highlight convergent results. Based on the retrospective assessment of baseline HRQoL after surgery and according to the LLRA, the trend of item easiness parameter is insignificant for social functioning. The corresponding MCA shows readjustment between response categories 1 and 2 and between response categories 3 and 4. In this way, as many patients had chosen higher than lower response categories at the retrospective measurement time as compared to the baseline measure, which is consistent with the LLRA. Based on retrospective assessment of baseline HRQoL at 3 months and according to the LLRA, patients had overestimated their body image with a positive trend of item easiness parameter. Regarding the corresponding MCA, it highlights a readjustment from response categories 2 to response categories 3 and from response categories 3 to response categories 4. In this way, patients had chosen higher response categories at the retrospective measurement time compared to the prospective measure indicating an overestimation of body image at baseline. Based on the retrospective assessment of the three months HRQoL at 6 months, patients had overestimated their pain level with a negative trend of item easiness parameter according to the LLRA. Regarding the MCA performed on pain at the same measurement times, it shows a recalibration between response categories 1 and 2, and only from response categories 4 to 3, not from 3 to 4.

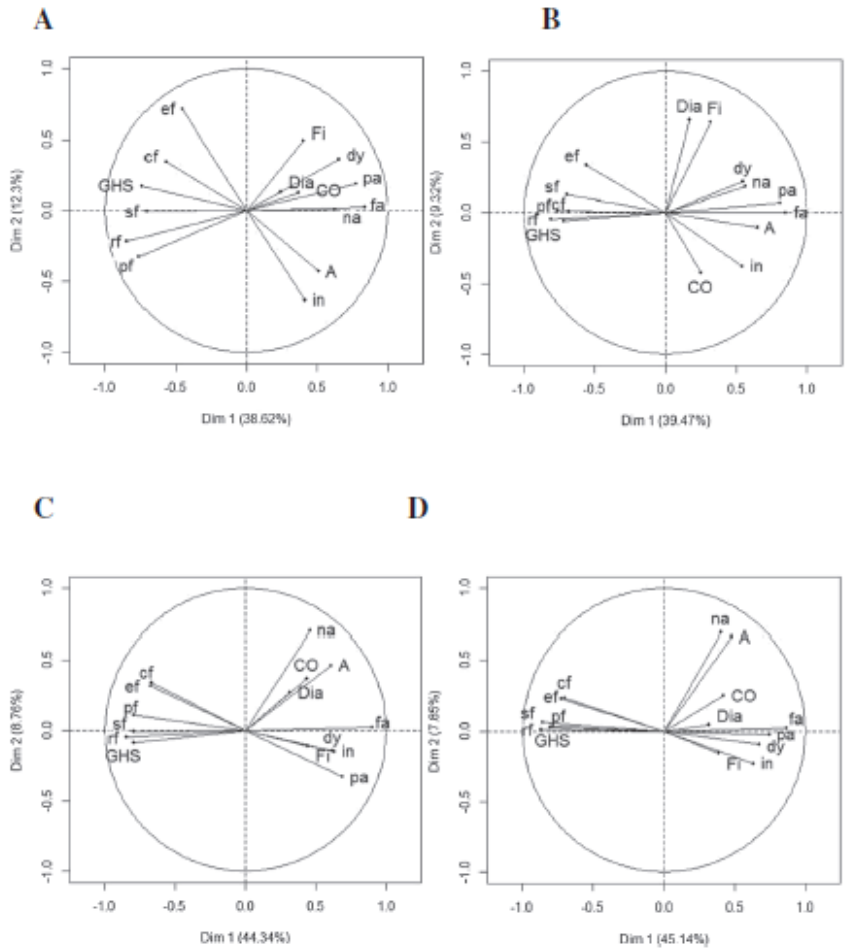


Figure 2 Graph representing the correlation between QLQ-C30 scores and the first two principal components of Principal Component Analysis at each prospective measurement time (N = 192): at baseline (Panel A), just after surgery (Panel B), at three months (Panel C) and at six months (Panel D). The QLQ-C30 measures five functional scales (physical functioning (pf), role functioning (rf), emotional functioning (ef), cognitive functioning (cf), social functioning (sf)), global health status (GHS), financial difficulties (Fi) and eight symptom scales (fatigue (fa), nausea and vomiting (na), pain (pa), dyspnea (dy), insomnia (in), appetite loss (A), constipation (CO), diarrhea (Dia)).

The secondary objective was to assess if PCA could be a valuable tool to longitudinally identify the reconceptualization and reprioritization components of RS independently of the occurrence of recalibration component of RS. PCA indicated a reprioritization of the HRQoL domains as evaluated by the QLQ-C30. Patients' anxiety probably related to the diagnosis of cancer and surgery seemed to be a major concern at baseline before the start of treatment, along with insomnia, which generated the second principal component, after fatigue and pain, which generated the first principal component. After surgery, diarrhea symptoms increased in importance, reflecting the impact of treatment. These results underline how patients adapt to their disease. At 3 months and 6 months, nausea and vomiting were more important as compared to diarrhea, also reflecting the toxicities of cancer treatment, especially chemotherapy. Regarding the QLQ-BR23, patients with a high level of systemic therapy side effects after surgery also tended to report a high level of arm symptoms, which can be due to the recent surgery. From 3 months, arm symptoms become less important, while future perspectives gained importance for primary breast cancer patients. Our results indicate that there is no correlation between breast symptoms and sexual functioning at 3 and 6 months.

No reprioritization was observed for the QLQ-C30 and QLQ-BR23 between the measures at M3 and M6. Patients seemed to assess their HRQoL with the same

Table 7 Correlation between quality of life scores and the first two first axis of principal component analysis on prospective measures

	T1: baseline		T2: after surgery		T3: 3 months		T4: six months	
Scores	First axis	Second axis	First axis	Second axis	First axis	Second axis	First axis	Second axis
QLQ-C30 (N = 192)								
GHS	−0.74	0.17	−0.72	−0.06	−0.79	−0.09	−0.79	0.03
Physical functioning	−0.76	−0.33	−0.76	0.01	−0.79	0.11	−0.81	0.03
Role functioning	−0.85	−0.22	−0.81	−0.05	−0.85	−0.05	−0.87	0.01
Emotional functioning	−0.46	0.71	−0.56	0.33	−0.68	0.31	0.73	0.23
Cognitive functioning	−0.57	0.34	−0.69	0.01	−0.67	0.34	−0.70	0.23
Social functioning	−0.71	−0.01	−0.70	0.13	−0.79	−0.01	−0.86	0.06
Fatigue	0.84	0.03	0.85	−0.01	0.90	0.02	0.87	0.02
Nausea and vomiting	0.63	0.01	0.56	0.18	0.46	0.70	0.40	0.69
Pain	0.77	0.19	0.81	0.07	0.69	−0.33	0.75	−0.02
Dyspnea	0.66	0.35	0.55	0.22	0.62	−0.14	0.67	−0.09
Insomnia	0.41	−0.63	0.54	−0.38	0.63	−0.15	0.63	−0.23
Appetite loss	0.51	−0.43	0.65	−0.11	0.61	0.46	0.48	0.66
Constipation	0.36	0.12	0.25	−0.42	0.44	0.37	0.42	0.25
Diarrhea	0.24	0.13	0.17	0.66	0.32	0.27	0.32	0.04
Financial difficulties	0.40	0.49	0.31	0.64	0.44	−0.11	0.38	−0.16
QLQ-BR23 (N = 154)								
Body image	−0.70	0.47	−0.64	0.45	−0.79	0.07	−0.79	0.25
Sexual functioning	0.30	−0.44	0.24	−0.52	0.31	−0.51	0.41	−0.41
Future perspective	−0.31	0.46	−0.53	0.52	−0.76	0.25	−0.73	0.30
Systemic therapy side effects	0.78	−0.01	0.72	0.28	0.61	−0.49	0.67	−0.21
Breast symptoms	0.58	0.49	0.64	0.38	0.56	0.66	0.57	0.64
Arm symptoms	0.66	0.49	0.75	0.32	0.69	0.44	0.63	0.57

relative importance at 6 months as they did at 3 months suggesting that after treatment initiation they have a more "stabilized" appreciation of HRQoL dimensions.

The reprioritization of symptomatic scales enables interpretation of HRQoL levels and changes and impact of treatments and disease on HRQoL. Then based on these results we suggest that the occurrence of the reprioritization component of RS should be taken into account in the interpretation of the results of the longitudinal analysis. Deterioration of a scale, which becomes more important over time for the patient, could have a strong impact on patient's overall HRQoL level and could indicate priority for care. Conversely, deterioration of a scale, which for the patient loses importance over time, could have a minor impact on patient's general HRQoL level.

Reconceptualization is reflected by changes in connections and contrasts between variables, and more generally by changes in graph structure of PCA. The functional scales of the QLQ-C30 became increasingly interrelated. When one functional scale is affected by cancer treatment or disease progression, then it is likely that all the other functional scales are affected. Moreover, patients had associated nausea and vomiting to appetite loss at 6 months.

These results suggest that PCA is an indirect method in investigating the reprioritization and reconceptualization components of RS.

The main limitation of this work is the use of the Then-test as the standard method to explore recalibration. The Then-test method is increasingly called into question [41-43], mainly because it can induce a recall bias [13]. Indeed, the second reassessment of baseline HRQoL was three months after baseline and the reassessment of HRQoL at M3 was three months after the prospective measure so it may induce a recall bias.

Schwartz et al. have proposed some guidelines to improve the stringency of the Then-test method [41]. In their paper, Schwartz et al. recommended to include a control group, which would not susceptible to RS. As RS is a treatment-dependent phenomenon, we tried to constitute a control group including patients with only a suspicion of BC. However, the number of patients with no confirmed BC was not sufficient to constitute

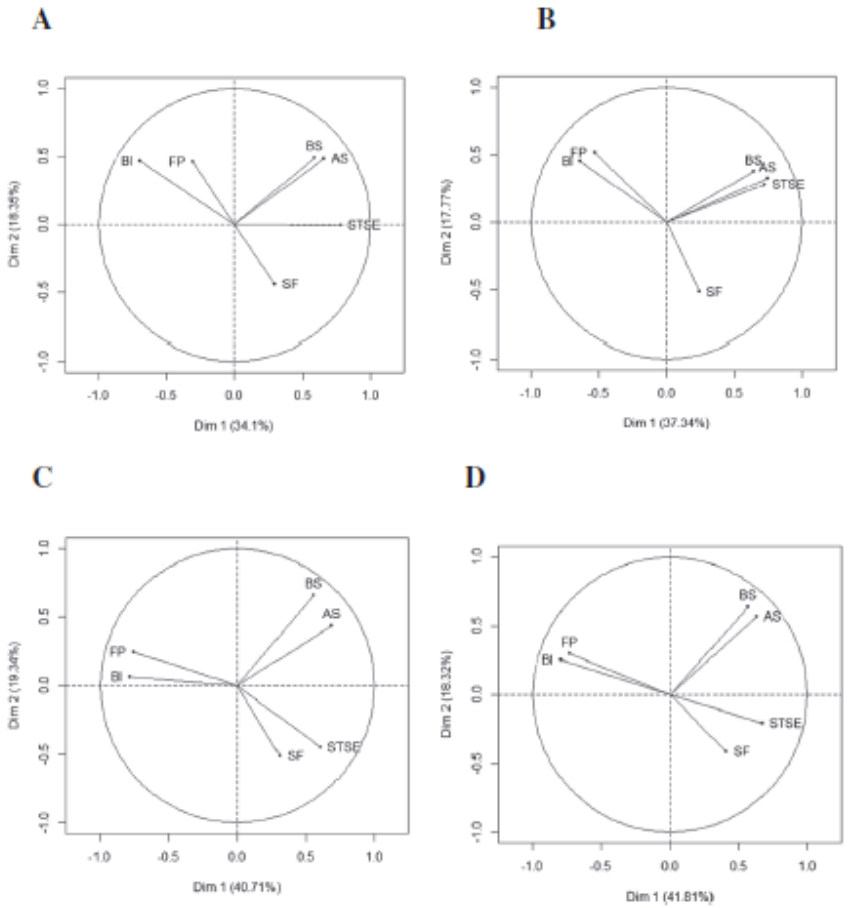


Figure 3 Graph representing the correlation between QLQ-BR23 scores and the first two principal components of Principal Component Analysis at each prospective measurement time (N = 154): at baseline (Panel A), just after surgery (Panel B), at three months (Panel C) and at six months (Panel D). The QLQ-BR23 measures four functional scales (body image (BI), sexual functioning (SEXF), sexual enjoyment (SE), future perspective (FP)) and four symptom scales (systemic therapy side effects (STSE), breast symptoms (BS), arm symptoms (AS), upset by hair loss (HL)). SE and HL are excluded from these analyses.

a control group. Others explanations of detection of RS effect could then also be formulated as social desirability responding. This hypothesis cannot be verified since a control group could not be constituted. As it was recommended by Schwartz et al., we reported effect sizes for recalibration in order to assess the magnitude of RS effect. A Bonferroni correction of type I error rate was performed in order to minimize false positive conclusions. The guidelines also recommended to use internal or external validation approaches as performance-based, perception-based, and evaluation-based items/subscales for internal validation of then-test results and clinical measures indicating health state at baseline and follow-up for external validation. However, we failed to include such approaches in this present study. The instructions of the retrospective questionnaires clearly indicate the patients to think back to the referent time as advisable by Schwartz et al. Moreover, the nomenclature used in this paper to characterize the recalibration component of RS is those recommended [41]. The Then-test method is based on the assumption that patients rate their HRQoL post-test and pre-test levels with the same criteria, since the assessments occur at the same time point. A test of the measurement invariance of the Then-test method would be necessary in this study in order to validate the then-test and to assess the possible recall bias due to its retrospective nature. This would be planned in another analysis using the Oort's procedure [16,44].

Based on this study, substituting the then-test with the LLRA and MCA to explore the recalibration component of RS cannot be recommended at this time. Nevertheless, IRT using LLRA could reinforce the Then-test method because of the improved interpretation of recalibration. This model is effective, and the results are clearer, more explicit

and easy to summarize and to interpret. These methods should be used in other studies to validate their ability to reinforce the then-test method.

SEM is often used nowadays to demonstrate RS [11,16,17,45-48]. These models are not dependent on the Then-test method. However, they are based on the raw score and not on the items. In this way, IRT as compared to SEM could be more informative. Moreover, at this time, SEM has never been applied to the EORTC HRQoL questionnaires in order to highlight occurrence of the response shift effect.

It would be interesting to compare the statistical method described in the present study (factor analysis and IRT) to SEM applied on prospective measure in another paper in order to check their ability to capture all the three components of RS. There is a need to investigate all these methods using simulated data in order to establish differences using these three methods.

Factor analysis presents the advantage of graphically exploring all the components of RS. This visual representation is interesting in order to explore reconceptualization, which is the most conceptual component of the RS effect. Moreover, at this time, few methods have been proposed to identify this component [16] and in our point of view no gold standard has emerged. In addition, no additional questionnaires are required for exploring reconceptualization and reprioritization. Thus, the use of PCA on the scores of the main questionnaires seems to be adequate in exploring these components. SEM is also often used to assess these components. However, our objective was to investigate the PCA method already used in the past [14,15,36] and not to apply SEM. Indeed, PCA is a special case of SEM.

IRT models and factor analysis are mostly used in the development and validation of HRQoL questionnaires [49-55]. However, several studies have begun to use IRT in longitudinal studies of HRQoL [29,56-61], underscoring the potential of these models in longitudinal analyses. Moreover, longitudinal IRT model was used in order to characterize recalibration component of RS. Few studies have investigated RS using IRT while differential item functioning based on IRT was also proposed as alternative approach [18,62,63].

Finally, PCA were performed on patients with all scores available at all the prospective measurement times. Only 40% to 50% of patients were thus retained in the analysis but these patients were comparable to those excluded according to baseline characteristics except there was an age effect which may reflect a selection bias.

The data presented in this article confirm the potential of IRT models in longitudinal HRQoL studies, especially their ability to characterize more precisely the recalibration component of RS. Our data also underline the interest of PCA to characterize reprioritization and reconceptualization components of RS. These results confirm the need to take recalibration into account when comparing longitudinal HRQoL data between patient groups and the need to explore the other components in order to better interpret results [64,65]. The items of these questionnaire are prone to response shift effect since they are evaluation-based items. Then an objective assessment by the patient cannot be made. Despite the fact that items of these questionnaires are prone to RS effect. Some work is still needed to provide both a longitudinal analysis method easy to understand for the clinician and to extract the potential measurement bias due to the occurrence of a response shift effect. Another solution would be to develop or use other questionnaires not prone to response shift effect with more performance-based items [41]. Future studies should investigate the ability of these statistical methods to capture all components of RS without the then-test method.

Abbreviations

BC: Breast cancer; GHS: Global Health Status; EORTC: European Organization for Research and Treatment of Cancer; HRQoL: Health-related quality of life; IRT: Item response theory; LLRA: Linear logistic model with relaxed assumptions; MCA: Multiple Correspondence Analyses; MCID: Minimal clinically important difference; MD: mean difference; PCA: Principal component analyses; RS: Response shift; SEM: Structural equation modelling; STSE: Systemic therapy side effects.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AA performed the statistical analyses and interpretation and written the manuscript, CBM interpreted the data and drafted the manuscript, TC, FG, MV, DJ, MM, SC, JC, OG designed the study, SC, JC, OG included the patients, ZH interpreted the data, FB designed the study, written protocol, managed the statistical analyses, interpreted the data and review the draft. All authors read and approved the final manuscript.

Acknowledgments

We thank Holly Sandu for correcting the manuscript. This work was supported by a grant from the "Institut National du Cancer". The study sponsor had no role in the conception, the design of the study, the data acquisition and analysis or in the manuscript preparation.

Author details

[1]Quality of Life in Oncology Platform, Besançon, France. [2]Methodological and Quality of Life Unit in Oncology, University Hospital of Besançon, Besançon, France. [3]EA 3181, University of Franche-Comte, Besançon, France. [4]Department of Biostatistics, Institut Régional du Cancer Montpellier, Montpellier, France. [5]Medical Oncology Department, Centre Alexis Vautrin, Nancy, France. [6]Clinical Epidemiology and Evaluation Department, Inserm, CIC-EC, and CHU, Nancy, France. [7]Department of Epidemiology and Public Health, Faculty of Medicine, EA 3430, University of Strasbourg, Strasbourg, France. [8]Pôle Recherche – Innovations, University Hospital of Reims, Reims, France. [9]Surgery Department, Centre Georges François Leclerc, Dijon, France. [10]Gynecological and Obstetric Department, Institut Mère Enfant, University Hospital of Reims, Reims, France. [11]Public health laboratory, EA 3279, Aix-Marseille University, Marseille, France.

Received: 27 January 2014 Accepted: 1 March 2014
Published: 8 March 2014

References

1. Osoba D: **Health-related quality of life and cancer clinical trials.** *Ther Adv Med Oncol* 2011, **3**:57–71.
2. Montazeri A: **Health-related quality of life in breast cancer patients: a bibliographic review of the literature from 1974 to 2007.** *J Exp Clin Cancer Res* 2008, **27**:32.
3. Ubel PA, Peeters Y, Smith D: **Abandoning the language of "response shift": a plea for conceptual clarity in distinguishing scale recalibration from true changes in quality of life.** *Qual Life Res* 2010, **19**:465–471.
4. Wiklund I: **Assessment of patient-reported outcomes in clinical trials: the example of health-related quality of life.** *Fundam Clin Pharmacol* 2004, **18**:351–363.
5. Bullinger M: **Assessing health related quality of life in medicine. An overview over concepts, methods and applications in international research.** *Restor Neurol Neurosci* 2002, **20**:93–101.
6. Gibbons FX: **Social comparison as a mediator of response shift.** *Soc Sci Med* 1999, **48**:1517–1530.
7. Howard GS, Dailey PR, Gulanick NA: **The feasibility of informed pretests in attenuating response shift bias.** *Appl Psychol Meas* 1979, **3**:481–494.
8. Howard GS, Ralph KM, Gulanick NA, Maxwell SE, Nance SW, Gerber SK: **Internal invalidity in pretest-posttest self-report evaluations and a reevaluation of retrospective pretests.** *Appl Psychol Meas* 1979, **3**:1–23.
9. Sprangers MA, Schwartz CE: **Integrating response shift into health related quality of life research: a theoretical model.** *Soc Sci Med* 1999, **48**:1507–1515.
10. Korfage IJ, de Koning HJ, Essink-Bot ML: **Response shift due to diagnosis and primary treatment of localized prostate cancer: a then-test and a vignette study.** *Qual Life Res* 2007, **16**:1627–1634.
11. Oort FJ, Visser MR, Sprangers MA: **An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery.** *Qual Life Res* 2005, **14**:599–609.
12. Schwartz CE, Sprangers MA: **Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research.** *Soc Sci Med* 1999, **48**:1531–1548.
13. Sprangers MA, Van Dam FS, Broersen J, Lodder L, Wever L, Visser MR, Oosterveld P, Smets EM: **Revealing response shift in longitudinal research on fatigue–the use of the thentest approach.** *Acta Oncol* 1999, **38**:709–718.
14. Schmitt N: **The use of analysis of covariance structures to assess beta and gamma change.** *Multivar Behav Res* 1982, **17**:343–358.
15. Ahmed S, Mayo NE, Corbiere M, Wood-Dauphinee S, Hanley J, Cohen R: **Change in quality of life of people with stroke over time: true change or response shift?** *Qual Life Res* 2005, **14**:611–627.
16. Oort FJ: **Using structural equation modeling to detect response shifts and true change.** *Qual Life Res* 2005, **14**:587–598.
17. King-Kallimanis BL, Oort FJ, Nolte S, Schwartz CE, Sprangers MA: **Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients.** *Qual Life Res* 2011, **20**:1527–1540.
18. Craig S: **Measuring Change Retrospectively: An Examination Based on Item Response Theory.** In *"Measuring Behavioral Change: Methodological Considerations." Symposium Presented at the Annual Conference of the Society for Industrial and Organizational Psychology.* Edited by Martineau J.; 2000.
19. Meade AW, Ellington JK, Craig SB: **Exploratory Measurement Invariance: A New Method Based on Item Response Theory.** In *Symposium Presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.*; 2004.
20. Dabakuyo TS, Guillemin F, Conroy T, Velten M, Jolly D, Mercier M, Causeret S, Cuisenier J, Graesslin O, Gauthier M, Bonnetain F: **Response shift effects on measuring post-operative quality of life among breast cancer patients: a multicenter cohort study.** *Qual Life Res* 2013, **22**:1–11.
21. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JC, Kaasa S, Klee M, Osoba D, Ravasi D, Rolbe PB, Schraub S, Sneeuw K, Sullivan M, Takeda F: **The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology.** *J Natl Cancer Inst* 1993, **85**:365–376.
22. Sprangers MA, Groenvold M, Arraras JI, Franklin J, te Velde A, Muller M, Franzini L, Williams A, de Haes HC, Hopwood P, Cull A, Aaronson NK: **The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: first results from a three-country field study.** *J Clin Oncol* 1996, **14**:2756–2768.
23. Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D: **Bottomley AobotEQoLG. In** *EORTC QLQ-C30 Scoring Manual.* 3rd edition. EORTC; 2001. edn; 2001.
24. Osoba D, Rodrigues G, Myles J, Zee B, Pater J: **Interpreting the significance of changes in health-related quality-of-life scores.** *J Clin Oncol* 1998, **16**:139–144.
25. Hoffman DL, De Leeuw J: **Interpreting multiple correspondence analysis as an multidimensional scaling method.** *Mark Lett* 1992, **3**:259–272.
26. Greenacre MJ: **Interpreting multiple correspondence analysis.** *Appl Stochastic Model Data Anal* 2006, **7**:195–210.
27. Fischer GH: **Some Probabilistic Models for Measuring Change.** In *Advances in Psychological and Educational Measurement.* Edited by de Gruijter DNM, van der Kamp LJT. New York: Wiley; 1976.
28. Fischer GH: **Some Latent Trait Models for Measuring Change in Qualitative Observations.** In *New Horizons in Testing. Latent Trait Theory and Computerized Adaptive Testing.* Edited by Weiss DJ. New York: Academic Press; 1983.
29. Fischer GH, Ponocny I: **An extension of the patial credit model with an application to the measurement of change.** *Psychometrika* 1994, **59**:177–192.
30. Fischer G, Parzer P: **An extension of the rating scale model with an application to the measurement of change.** *Psychometrika* 1991, **56**:637–651.
31. Rasch G: *Probabilistic Models for Some Intelligence and Attainment Tests.* The Danish Institute of Educational Research, Copenhagen: MESA Press; 1960.
32. Fischer GH: *Linear Logistic Models for Change. Rasch Models. Foundations, Recent Developments and Applications.* New York: Springer; 1995.
33. Van der Linden WJ, Hambleton RK: *Handbook of Modern Item Response Theory.* New York: Springer Verlag; 1997.
34. Mair P, Hatzinger R: **CML based estimation of extended Rasch models with the eRm package in R.** *Psychol Sci Q* 2007, **49**:26–43.
35. Mair P, Hatzinger R: **Extended rasch modeling: the eRm package for the application of IRT models in R.** *J Stat Softw* 2007, **20**:1–20.
36. Barclay-Goddard R, Epstein JD, Mayo NE: **Response shift: a brief overview and proposed research priorities.** *Qual Life Res* 2009, **18**:335–346.
37. Costello AB, Osborne JW: **Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis.** *Pract Asses Res Eval* 2005, **10**:173–178.
38. Development Core Team R: **R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.** URL http://www.R-project.org/.
39. Lê S, Josse J, Husson F: **FactoMineR: an R package for multivariate analysis.** *J Stat Softw* 2008, **25**:1–18.
40. Hatzinger R, Rusch T: **IRT models with relaxed assumptions in eRm: a manual-like instruction.** *Psychol Sci Q* 2009, **51**:87–120.
41. Schwartz CE, Sprangers MA: **Guidelines for improving the stringency of response shift research using the thentest.** *Qual Life Res* 2010, **19**:455–464.
42. Schwartz CE, Rapkin BA: **Understanding appraisal processes underlying the thentest: a mixed methods investigation.** *Qual Life Res* 2012, **21**:381–388.
43. Visser MR, Oort FJ, Sprangers MA: **Methods to detect response shift in quality of life data: a convergent validity study.** *Qual Life Res* 2005, **14**:629–639.
44. Nolte S, Elsworth GR, Sinclair AJ, Osborne RH: **Tests of measurement invariance failed to support the application of the "then-test".** *J Clin Epidemiol* 2009, **62**:1173–1180.
45. Ahmed S, Bourbeau J, Maltais F, Mansour A: **The Oort structural equation modeling approach detected a response shift after a COPD self-management program not detected by the Schmitt technique.** *J Clin Epidemiol* 2009, **62**:1165–1172.
46. Gandhi PK, Ried LD, Huang IC, Kimberlin CL, Kauf TL: **Assessment of response shift using two structural equation modeling techniques.** *Qual Life Res* 2013, **22**:461–471.
47. Donaldson GW: **Structural equation models for quality of life response shifts: promises and pitfalls.** *Qual Life Res* 2005, **14**:2345–2351.
48. King-Kallimanis BL, Oort FJ, Visser MR, Sprangers MA: **Structural equation modeling of health-related quality-of-life data illustrates the measurement and conceptual perspectives on response shift.** *J Clin Epidemiol* 2009, **62**:1157–1164.

49. Edelen MO, Reeve BB: Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res* 2007, 16(Suppl 1):5–18.
50. Floyd FJ, Widaman KF: Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assess* 1995, 7:286–299.
51. Hambleton RK: Emergence of item response modeling in instrument development and data analysis. *Med Care* 2000, 38:II60–II65.
52. Lai JS, Crane PK, Cella D: Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. *Qual Life Res* 2006, 15:1179–1190.
53. McLachlan SA, Devins GM, Goodwin PJ: Factor analysis of the psychosocial items of the EORTC QLQ-C30 in metastatic breast cancer patients participating in a psychosocial intervention study. *Qual Life Res* 1999, 8:311–317.
54. Reise SP, Widaman KF, Pugh RH: Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol Bull* 1993, 114:552–566.
55. Smith AB, Wright P, Selby PJ, Velikova G: A rasch and factor analysis of the functional assessment of cancer therapy-general (FACT-G). *Health Qual Life Outcomes* 2007, 5:19.
56. Liu L, Hedeker D, Mermelstein R: Modeling nicotine dependence: an application of a longitudinal IRT model for the analysis of adolescent nicotine dependence syndrome scale. *Nicotine Tob Res* 2013, 15:326–333.
57. Blanchin M, Hardouin JB, Le Neel T, Kubis G, Blanchard C, Mirallié E, Sébille V: Comparaison of CTT and rasch-based approaches for the analysis of longitudinal patient reported outcomes. *Stat Med* 2011, 30:825–838.
58. Swartz RJ, Schwartz C, Basch E, Cai L, Fairclough DL, McLeod L, Mendoza TR, Rapkin B: The king's foot of patient-reported outcomes: current practices and new developments for the measurement of change. *Qual Life Res* 2011, 20:1159–1167.
59. van Nispen RM, Knol DL, Neve HJ, van Rens GH: A multilevel item response theory model was investigated for longitudinal vision-related quality-of-life data. *J Clin Epidemiol* 2010, 63:321–330.
60. Douglas JA: Item response models for longitudinal quality of life data in clinical trials. *Stat Med* 1999, 18:2917–2931.
61. Glas CA, Geerlings H, van de Laar MA, Taal E: Analysis of longitudinal randomized clinical trials using item response models. *Contemp Clin Trials* 2009, 30:158–170.
62. Craig S: Implicit theories and beta change in longitudinal evaluations of training effectiveness: an investigation using item response theory. 2002. Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.
63. Schwartz C, Martha J, Swaim B, Bode R, Kim D: Detecting response shift in using rasch analysis of then test data. *Qual Life Res* 2007, A-15(suppl).
64. Bonnetain F, Dahan L, Maillard E, Ychou M, Mitry E, Hammel P, Legoux JL, Rougier P, Bedenne L, Seitz JF: Time until definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. *Eur J Cancer* 2010, 46:2753–2762.
65. Hamidou Z, Dabakuyo TS, Mercier M, Fraisse J, Causeret S, Tixier H, Padeano MM, Loustalot C, Culsenier J, Sauzedde JM, Small M, Combier JP, Chevillote P, Rosbuger C, Arveux P, Bonnetain F: Time to deterioration in quality of life score as a modality of longitudinal analysis in patients with breast cancer. *Oncologist* 2011, 16:1458–1468.

doi:10.1186/1477-7525-12-32
Cite this article as: Anota *et al.*: Item response theory and factor analysis as a mean to characterize occurrence of response shift in a longitudinal quality of life study in breast cancer patients. *Health and Quality of Life Outcomes* 2014 12:32.

4.2. Modèles à équations structurelles

Résumé

Contexte

La qualité de vie relative à la santé (QdV) est un concept dynamique dépendant de l'adaptation du patient à sa maladie et reflété par un effet "Response Shift" (RS). Cet effet résulte d'un changement dans les références internes du patient (« recalibration »), dans l'importance relative des dimensions de QdV (« reprioritization ») et dans la conceptualisation de la QdV (« reconceptualization »). L'analyse longitudinale de la QdV doit tenir compte de l'occurrence éventuelle de cet effet. Les modèles à équations structurelles (« Structural Equation Modeling », SEM) sont une approche statistique proposée pour sa caractérisation mais à ce jour cette méthode n'est généralement appliquée qu'au questionnaire générique SF-36 respectant une structure particulière en deux macros domaines (physique et mental). L'objectif est d'investiguer ces modèles pour caractériser l'occurrence de la Response Shift sur deux questionnaires de l'EORTC spécifiques du cancer auprès de femmes atteintes d'un cancer du sein.

Méthodes

Une cohorte prospective multicentrique a été menée incluant toute femme atteinte d'un cancer du sein primitif. La QdV a été évaluée par les questionnaires EORTC QLQ-C30 et son module spécifique du cancer du sein BR23 à l'inclusion, après la première hospitalisation, à 3 et 6 mois. Différents modèles de mesure ont été explorés. Le modèle final intègre à la fois les échelles fonctionnelles et symptomatiques des deux questionnaires et considère que l'état symptomatique de la patiente impacte son état fonctionnel. La procédure de Oort a été appliquée afin de mettre en évidence chaque composante de la Response shift (Oort, 2005). La reconceptualisation n'a pas été explorée compte tenu de la construction du modèle de mesure. La procédure a été appliquée en intégrant les mesures de QdV 1) à l'inclusion et après la première hospitalisation, 2) à l'inclusion et à 3 mois et 3) à 3 mois et à 6 mois.

Résultats

381 patientes ont été incluses entre février 2006 et février 2008 dont 89% avaient un cancer du sein confirmé. Les analyses mettent en évidence un effet Response Shift à chaque temps de

mesure. L'indice RMSEA d'ajustement du modèle final varie entre 0,101 et 0,125. Après la première hospitalisation et à 3 mois, les analyses montrent :

- une recalibration non uniforme du fonctionnement de rôle, émotionnel et social, de l'image corporelle et des symptômes au niveau du bras par rapport à l'inclusion ;
- une recalibration uniforme du fonctionnement émotionnel et cognitive ainsi que des symptômes au niveau du sein avec une augmentation de l'intercept caractérisant une sous-estimation de ces échelles à l'inclusion ;
- et une augmentation de l'importance relative de la dimension du fonctionnement de rôle.

Peu d'effet Response shift est démontré à 6 mois comparativement à la mesure effectuée à 3 mois.

Conclusion

Les modèles SEM ont permis de mettre en évidence les composantes recalibration et changements de l'importance relative des domaines de QdV de la Response shift sur les questionnaires EORTC sans la nécessité d'une mesure rétrospective telle que l'approche then-test. Cependant, les données présentent un mauvais ajustement au modèle ce qui suggère que les modèles SEM ne seraient pas adaptés à la structure des questionnaires de l'EORTC.

<u>Article:</u> Is Structural Equation Modeling a valuable tool to characterize the occurrence of the Response Shift Effect with EORTC health-related quality of life questionnaires?

Article en cours de rédaction

Amélie Anota[1,2], Caroline Bascoul-Mollevi[3], Thierry Conroy[4], Francis Guillemin[1,5], Michel Velten[1,6], Damien Jolly[1,7], Zeinab Hamidou[1,8], Franck Bonnetain[1,2]

[1] National Platform Quality of Life and Cancer, France

[2] Methodological and Quality of Life Unit in Oncology (EA 3181), University Hospital of Besançon, Besançon, France

[3] Biostatistics Unit, ICM – Val d'Aurelle, Montpellier, France

[4] Medical Oncology Department, Centre Alexis Vautrin, Nancy, France

[5] Clinical Epidemiology and Evaluation Department, Inserm, CIC-EC, and CHU, Nancy, France

[6] Department of Epidemiology and Public Health, Faculty of Medicine, EA 3430, University of Strasbourg, Strasbourg, France

[7] Pôle Recherche – Innovations, University Hospital of Reims, Reims, France

[8] Public health laboratory, EA 3279, Aix-Marseille University, Marseille, France

Corresponding author:

Amélie Anota

Quality of Life and Cancer National Platform

Methodological and Quality of Life in Oncology Unit (EA 3181)

University Hospital of Besançon

France

Email: aanota@chu-besancon.fr

Telephone number: +33381218896

ABSTRACT

Background

Health-related quality of life is a dynamic concept reflected by the occurrence of a Response Shift (RS) effect. RS results in recalibration, reprioritization and reconceptualization of key HRQoL domains. Structural Equation Modeling (SEM) is a statistical approach often used to characterize the occurrence of the RS effect. However, this method is generally applied on SF-36 generic questionnaire. The objective was to investigate SEM to characterize RS effect on two EORTC HRQoL cancer specific questionnaires among women with breast cancer.

Methods

A prospective multicenter study including all primary breast cancer patients or suspicion was done. HRQoL was evaluated using the EORTC QLQ-C30 and QLQ-BR23 breast cancer module at baseline (T_0), after the first hospitalization (T_1), at 3 (T_2) and 6 months (T_3). The Oort procedure was used to characterize the Response Shift effect. The measurement model integrates both functional and symptomatic scales of both questionnaires and considers patient's symptomatic status impacts her functional status. Reconceptualization was not explored considering the construction of the measurement model. The procedure was applied integrating 1) baseline and after the first hospitalization measures, 2) baseline and 3 months measures and 3) 3 and 6 months measures.

Results

Between 2006 and 2008, 381 patients were included, 90% had a confirmed breast cancer. Uni-items dimensions were excluded of the final measurement model. At each measurement time, analyses highlighted the occurrence of a RS effect but with a poor fit to the final model (RMSEA between 0.101 and 0.125). At T_1 and T_2, a non uniform recalibration was highlighted for 7 and 9 dimensions respectively among the 13 dimensions retained, a uniform recalibration for 6 dimensions and a reprioritization for 4 dimensions. Few RS effect occurred at T_3.

Conclusions

SEM allows to highlight recalibration and reprioritization components of RS on EORTC questionnaires without the necessity of a retrospective measure such a then-test. Although these models are often used with SF-36 generic questionnaire, they seem to be not the most appropriated statistical approach to the structure of the EORTC HRQoL questionnaires.

Keywords: Health-related quality of life, Response Shift, Structural Equation Modeling, EORTC questionnaires

Abbreviations:

CFI: Comparative Fit Index

DIF: differential item functioning

GHS: Global Health Status

HRQoL: Health-Related Quality of Life

IRT: Item Response Theory

MCID: Minimal Clinically Important Difference

RMSEA: Root Mean Square Error of Approximation

RS: Response Shift

SD: Standard Deviation

SEM: Structural Equation Modeling

STSE: Systemic Therapy Side Effects

INTRODUCTION

Health-related quality of life (HRQoL) is an important endpoint in oncology [1]. Most of clinical trials must integrate HRQoL as an endpoint in order to investigate the clinical benefit for the patient.

HRQoL is generally assessed at several time points in oncology clinical trials: at baseline (before randomization), one or several times during the treatment administration and at the end of the study. Thus, the longitudinal analysis of these HRQoL data aims to determine the extent to which treatment toxicities or disease progression can affect patients' HRQoL level.

However, HRQoL is a subjective endpoint since it depends on patient's perception of his health status, his health expectations and definition of HRQoL. Patients can adapt to disease and treatment toxicities and then their health and HRQoL expectations can also change over time. These changes could result in a response shift (RS) effect.

Sprangers and Schwartz defined RS as "a change in the meaning of one's self-evaluation of a target construct as a result of: (a) a change in the respondent's internal standards of measurement (i.e. scale recalibration); (b) a change in the respondent's values (i.e. the importance of component domains constituting the target construct, [reprioritization]) or (c) a redefinition of the target construct (i.e. reconceptualization)" [2].

The RS effect is now a well-established concept and has been detected in various situation and cancer sites as in early breast cancer patients [3, 4]. Different approaches exist to highlight the occurrence of the RS effect [5, 6]. The most well-known method is the Then-test method which consist to introduce in the study design a retrospective measure of HRQoL [5]. Several statistical methods have also been proposed to characterize the occurrence of the RS effect [7-9].

In 2005, Oort proposed a sequential procedure based on structural equation modeling [7]. This method is attractive because it can highlight the three components of RS effect. Moreover, the procedure is relatively easy to replicate with an appropriate software. The Oort procedure was thus applied in several studies and to date, it is the most used method to highlight RS effect [10-12].

However, this procedure is generally applied to the SF-36 questionnaire. The SF-36 is a generic questionnaire respecting a particular structure in two macro domains (mental health and physical health) [13]. In oncology, some cancer specific questionnaires are preferred to generic ones because they are more adapted to the cancer patients and they offer a high responsiveness to change [14].

The European Organization of Research and Treatment of Cancer (EORTC) has developed a core questionnaire QLQ-C30 specific to cancer [15]. According to the cancer sites or treatment modality, supplementary modules have been developed to use in conjunction with the QLQ-C30 such as the QLQ-BR23 module for breast cancer [16]. The QLQ-C30 questionnaire as well as all its supplementary module is a multi-dimensional questionnaire, with multi- or uni-item dimensions, corresponding to functional and symptomatic scales. Contrary to the SF-36, the HRQoL dimensions assessed by the QLQ-C30 are not combined in one or more macro domains or HRQoL components.

Some previous studies have already investigated the SEM method on the QLQ-C30 questionnaire and QLQ-BR23 module but only to test structural invariance among patients groups [17, 18]. Moreover, these studies retained few dimensions of these questionnaires. To our knowledge, the Oort procedure to test the occurrence of the RS effect has never been applied at this time on the EORTC HRQoL questionnaires.

Moreover, the SEM are generally performed with two measurement time points while in oncology clinical trials more than two assessments is usually planned.

The aim of this study was thus to investigate the Oort procedure on two cancer specific EORTC HRQoL questionnaires to characterize the occurrence of the RS effect for HRQoL in breast cancer patients at several measurement time points.

METHODS

Patients and eligibility criteria

A prospective, multicenter, randomized cohort study was performed in four French centers. All women initially hospitalized between February 2006 and February 2008 for diagnosis or treatment of primary or suspected breast cancer were eligible for inclusion. Women with cancer other than breast cancer, already undergoing breast cancer treatment, or with a previous history of cancer were excluded. Written informed consent was obtained from every participant and the protocol was approved by Ethics committee. The complete design of this study was extensively described elsewhere [3].

Health-related Quality of Life Assessment

HRQoL was evaluated using the EORTC QLQ-C30 cancer specific questionnaire [15] and its breast cancer module QLQ-BR23 [16] at four time points: at diagnosis (T_0), at the end of the initial hospitalization (T_1), at three (T_2) and six months (T_3) after the first hospitalization.

The QLQ-C30 questionnaire contains 30 items allowing to measure five functional scales (physical, role, emotional, cognitive and social functioning), global health status (GHS), financial difficulties and eight symptom scales (fatigue, nausea and vomiting, pain, dyspnea, insomnia, appetite loss, constipation, diarrhea) [15].

The QLQ-BR23 module contains 23 items assessing four functional scales (body image, sexual functioning, sexual enjoyment, future perspective) and four symptom scales (systemic therapy side effects (STSE), breast symptoms, arm symptoms, upset by hair loss) specific to breast cancer [16].

Scores were calculated if at least half of the items were answered according to the recommendations of the EORTC Scoring Manual [19]. These scores vary from 0 (worst) to 100 (best) for the functional dimensions and GHS, and from 0 (best) to 100 (worst) for the symptom dimensions.

Descriptive analysis

Baseline variables were described using means and standard deviations (SD) for continuous variables and percentages for qualitative variables. The number of HRQoL questionnaires completed at each measurement time was reported.

Response shift detection according to the Oort procedure

The SEM analysis investigating RS was performed integrating two measurement times:

1) At baseline (T_0) and after surgery (T_1)
2) At baseline (T_0) and at three months (T_2)
3) At three months (T_2) and three months (T_3).

For each analysis, all patients with at least one available HRQoL at one measurement time were retained (modified intent-to treat analysis).

Step 1: Establishing a measurement model

We aimed to find a satisfactory measurement model integrating both the EORTC QLQ-C30 and QLQ-BR23 scales. Sexual functioning, sexual enjoyment and upset by hair loss dimensions were excluded due to the occurrence of a lot of missing data. Several measurement models were investigating:

- included or not uni-item dimensions;
- considering both functional and symptomatic scales in the same model or separately.

Regarding the model with both functional and symptomatic, we considered that symptomatic state of the patient impacts it functional states and HRQoL level according to the Wilson and Cleary model [20]. Thus, two latent components were considered: functional and symptomatic latent variables.

In SEM analysis, a HRQoL score observed x_i is linked to a given latent variable X by the following equation:

$$x_i = \alpha_i + \beta_i X + e.x_i$$

with α_i is the intercept, β_i is the factor loading and $e.x_i$ is the residual variance.

According to the Oort procedure a change between two measurement times in:

- the residual variance corresponds to the occurrence of the non uniform recalibration component of the RS effect;
- the intercept correspond to the occurrence of the uniform recalibration;
- the factor loading corresponds to the reprioritization component of the RS effect.

Given the construction of the measurement model, the reconceptualization component of the RS effect could not be explored.
A distinction is made between “uniform” and “non uniform” recalibration for RS effect as for Differential Item Functioning analysis. The recalibration is qualified as “uniform” if the change affects all response categories in the same direction and in the same extent. Conversely, a “non uniform” recalibration” represents a distortion of the measurement scale (i.e. stretch or shrink).

To assess the overall goodness-of-fit of our models, the root mean square error of approximation (RMSEA) [21] and the comparative fit index (CFI) were considered [22]. An RMSEA value of <0.08 indicates satisfactory fit and a value of <0.05 indicates close fit [21]. The CFI assess the improvement in fit from a null model that assumes no relationships between variables. Values of >0.95 for the CFI indicate reasonable fit of the model to data. Models were estimated according to maximum likelihood with missing data method of STATA software.

A model was then constructed integrating two measurement times. In this model (Model 1), a RS effect was allowed. Conversely, no true change was allowed. In this way, the mean and variance of both measurement times are fixed to 0 and 1 respectively. No constraint was imposed associated to RS effect.

Step 2: Overall test of response shift

In this step, a second model was fitted in which no RS effect was allowed (Model 2). In this model, all constraints of the invariance associated to RS effect were imposed and true change was allowed. To allow the identifiability of the model, the mean and variance of the latent variable at the first measurement time were fixed to 0 and 1 respectively.

An overall test was then performed comparing the fit of the Model 1 and Model 2 using the likelihood ratio test [23].

If this test was not statistically significant, no RS effect was suspected and the next step was step 4.

Step 3: Detection of Response Shift effects

In this step, the RS effect was detected according to the step by step procedure of Nolte et al. [24]. In this way, a Model 3 was constructed from the Model 2 and each invariance constraint associated to RS was removed in a sequential order in order to detect first the non-uniform recalibration, then the uniform recalibration and finally the reprioritization component of the RS effect. The objective was to improve the fit of the Model 2. At each step, the most significant RS effect detected according to the likelihood ratio test was integrated in the new Model 2. Only effects which were then test again.

All tests were two-sided and the type I error was set to 0.05. No adjustment for multiple testing was done since it was a secondary endpoint of the trial. All analyses were performed with SE 13.1 software (Stata Corporation, College Station, TX, USA).

RESULTS

Population and HRQoL completion

381 patients were included between February 2006 and February 2008. Mean age was 58.4 (SD=11) years. Three hundred and forty (89%) patients had confirmed breast cancer. Baseline characteristics of the patients are summarized in Table 1.

Median time between baseline and the end of the first hospitalization was 6 days.

At baseline, 359 (94%) patients had at least one HRQoL score available, 343 (90%), 340 (89%) and 321 (84%) at the end of the first hospitalization (i.e. after surgery), and at three and six months respectively.

Response shift detection according to the Oort procedure

Final Measurement model

The most satisfactory measurement model integrated dimensions of both the QLQ-C30 and QLQ-BR23 questionnaires, considered both functional and symptomatic scales in the same model and excluded uni-item dimensions. The corresponding path diagram was reported in Figure 1.

Response Shift detection just after surgery

First, the occurrence of the RS effect was investigated just after the first hospitalization (T_1) as compared to the baseline measure at diagnosis (T_0). 377 patients (98.9%) were retained in this analysis with at least one HRQoL score available at one measurement time.

The comparison of Model 1 with RS effect and Model 2 with no RS effect indicated that a RS effect occurred with a P-value <0.0001 for the likelihood ratio test (Table 2).

Thus, the step by step procedure of Nolte et al. [24] was applied. Both recalibration and reprioritization were highlighted in the final model (Table 3).

A non uniform recalibration for 7 dimensions among the 13 retained: physical, role, emotional, and social functioning, body image, fatigue and arm symptoms. The residual variance increases for all these dimensions excepted emotional functioning, qualifying an enlargement of the measurement scale.

A uniform recalibration was highlighted for 6/13 dimensions:

- for emotional and cognitive functioning with an increase of the intercept from 64.9 to 80.5 and from 83.5 to 90.6 respectively. Thus, patients underestimated their emotional and cognitive level at baseline;
- for fatigue and STSE with a decrease of the intercept from 22.6 to 19.5 and from 12.9 to 8.2 respectively. Thus, patients overestimated the presence of these symptoms at baseline;
- and for arms and breast symptoms with an increase of the intercept from 7.7 to 10.0 and from 11.0 to 15.5 respectively. Thus, patients underestimated the presence of these symptoms at baseline.

Finally, a reprioritization was highlighted for 4/13 dimensions:

- role functioning and “nausea and vomiting” were less important just after surgery as compared to baseline measure;
- STSE and arm symptoms were more important just after surgery as compared to baseline measure.

However, a poor model fit was observed (RMSEA=0.106, Table 2).

Response detection at three months

The occurrence of the RS effect was then investigated at three months after the first hospitalization (T_2) as compared to the baseline measure (T_0). 378 patients (99.2%) were retained in this analysis with at least one HRQoL score available at one measurement time.

The comparison of Model 1 with RS effect and Model 2 with no RS effect indicated that a RS effect occurred with a P-value <0.0001 for the likelihood ratio test (Table 2).

A non uniform recalibration for 9/13 dimensions (Table 4): role, emotional, and social functioning, body image, pain, nausea and vomiting, STSE, breast and arm symptoms with an increase of the residual variance for all these dimensions except emotional functioning.

A uniform recalibration was highlighted for 6/13 dimensions.

With an increase of the intercept, patients underestimated their GHS (68.8 to 72.3), emotional (64.8 to 83.6) and cognitive functioning (83.4 to 89.3) and their breast symptoms level (11.1 to 17.0) at baseline as compared to 3 months after surgery.

With a decrease of the intercept, patients overestimated their role functioning (89.7 to 88.3) and body image (89.6 to 83.3) at baseline as compared to the HRQoL measure at 3 months.

A reprioritization was also highlighted for 4/13 dimensions with an increase of the factor loading at three months as compared to baseline: role (1.6 to 2.0) and social functioning (1.3 to 1.8), body image (0.9 to 1.3) and breast symptoms (5.5 to 8.4) were more important for the patients at three months as compared to baseline.
However, a poor model fit was observed (RMSEA=0.106, Table 2)

Response shift detection at six months

Finally, the occurrence of the RS effect was investigated at six months after surgery (T_3) as compared to the three months measure (T_2). 349 patients (91.6%) were retained in this analysis with at least one HRQoL score available at one measurement time.

The comparison of Model 1 with RS effect and Model 2 with no RS effect indicated that a RS effect occurred with a P-value =0.0002 for the likelihood ratio test (Table 2).

A non uniform recalibration for 3/13 dimensions (Table 5): social functioning, nausea and vomiting and breast symptoms. A decrease of the residual variance was observed for these dimensions qualifying a shrinkage of the measurement scale at 6 months compared to at 3 months.

A uniform recalibration was highlighted for 4/13 dimensions. Patients underestimated their GHS, pain and arm symptoms levels at three months as compared to six months after surgery when they evaluated their HRQoL level.

No reprioritization was highlighted in this model and the final model presented a very poor model fit (RMSEA=0.125, Table 2)

DISCUSSION

This study was the first one investigating the Oort procedure [7] to characterize the occurrence of RS effect in HRQoL study using EORTC questionnaires. Moreover, a longitudinal assessment of RS was performed investigating the occurrence of the RS effect at several measurement times: just after surgery, three and six months after surgery.

In this study, the SEM allowed to highlight both the recalibration (uniform and non uniform) and reprioritization components of the RS effect. Most of the RS effects occurred at three months after the initial hospitalization. Few RS effect occurred at 6 months compared to at three months.

After the initial hospitalization and at three months, an increase of the intercept was observed for emotional and cognitive functioning and for breast symptoms. For a given HRQoL level, patients reported higher scores for these dimensions after surgery and at three months than they did at baseline. At these time points, a non-uniform recalibration was observed for role, emotional and social functioning, body image and arm symptoms. Moreover, the role functioning was more important for patients regarding their global HRQoL level than at baseline (factor loading increased from 1.6 to 2.0).

After the initial hospitalization, the fatigue and STSE intercepts decreased as compared to the baseline measure. Thus, considering the HRQoL measure after the initial hospitalization as the new internal reference for the patients, patients overestimated the presence of these symptoms and side effects at baseline. The presence of arm symptoms had a higher impact on patient's HRQoL level after surgery than at baseline (factor loading increased from 8.6 to 11.2). Moreover, considering the HRQoL measure after the initial hospitalization as the new internal reference for the patients, patients underestimated their arms symptoms level at baseline as compared to after the initial hospitalization.

At three months, social functioning, body image and breast symptoms were more important for the patients regarding their global HRQoL. No reprioritization was observed at six months as compared to at three months, suggesting that the importance of these dimensions remained unchanged.

The SEM analyses have shown several limitations in this study. An important limitation is that uni-item dimensions were excluded of these analyses. In fact, using the maximum likelihood estimation method, the SEM required a normally distribution of the observed scores. Uni-item dimensions more reflect an ordinal scale than a normally scale.

The reconceptualization component of the RS effect could not be investigated in this study due to the construction of the measurement model. Two latent components were considered: the symptomatic the functional states of the patient. Indeed, we considered that the symptomatic state impacts the functional state of the patients. To be able to explore the reconceptualization, HRQoL scores should be split in two or more HRQoL domains. For example, the SF-36 is a questionnaire which HRQoL dimensions are separated in two domains: mental health and physical health [13]. As the Confirmatory Factor Analysis cleary highlights these two domains for SF-36 questionnaire, there is no agreed upon the structure for the QLQ-C30 [17].

Finally, models retained presented a poor model fit with an RMSEA between 0.101 and 0.126. An acceptable model fit must present an RMSEA value of less than 0.08.

The SEM analysis through Oort procedure is an attractive statistical method since it can highlight all the types of RS effect. Moreover, this method only used the prospective HRQoL measures. However, they have shown several limitations in this study indicating that SEM seems to be not well adapted to the structure of the EORTC HRQoL questionnaires. Uni-item dimensions could not be considered in SEM analysis since normally distributions of the scales could not be respected. Most of the symptoms scales of the EORTC QLQ-C30 questionnaire are uni-item dimensions. Six over 15 dimensions of the QLQ-C30 are uni-item dimensions. These dimensions assess important symptoms and treatment side effects experienced by the patients during the treatment. A RS effect could affect these dimensions and thus they have to be taken into account in the RS detection analysis.

To characterize the occurrence of the RS effect, an alternative method could by to use Item Response Theory (IRT) [23]. These models could be more relevant for EORTC questionnaires than SEM since they are more adapted to ordinal scales. IRT models are focused on items themselves and not on the observed scores. At this time, IRT models are yet unexploited to characterize the RS effect while they are

widely used to validate the questionnaire psychometric properties and to test the presence of Differential Item Functioning[25].

Most of IRT models require the unidimensionality of the scale [26] while HRQoL is a multidimensional concept and EORTC questionnaire are multidimensional tools. Moreover, items of the EORTC questionnaire are constructed on a 4-point Likert scale (polytomous items). Thus, a multidimensional IRT model adapted to polytomous items [27] have to be investigate to characterize the Response Shit effect with the EORTC HRQoL questionnaires.

One limitation of the SEM (whatever the questionnaire used) is their inability to report clinically significant results. Results can be statistically significant but not clinically meaningful for the patient. It seems essential to take into the minimal clinically important difference and particularly in the quantification of the recalibration effect since it directly affects the assessment of the patient's HRQoL level on a given measurement scale.

Moreover, although several statistical approaches can detect the occurrence of the RS effect, no recommendation was given on how to take into account the occurrence of the RS effect in the longitudinal analysis of HRQoL data. Further investigations are thus still needed in this field in order to performed robust longitudinal analysis of HRQoL data.

To conclude, further investigations of statistical methods to characterize the occurrence of RS effect for HRQoL studies using EORTC HRQoL questionnaire are still needed. These statistical methods have to be adapted to the structure of theses questionnaires and must translate findings into clinical meaningful results for both clinicians and patients.

REFERENCES

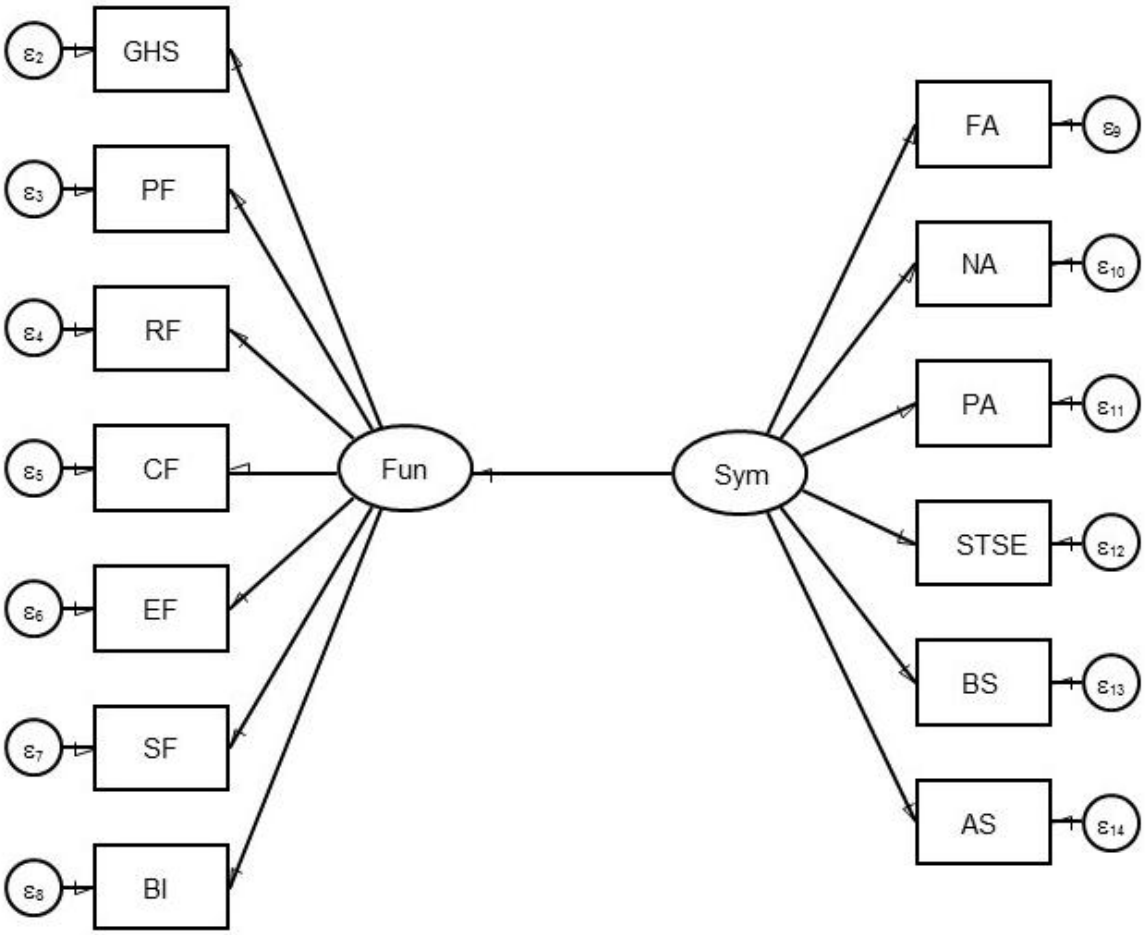
1. Osoba, D., *Health-related quality of life and cancer clinical trials.* Ther Adv Med Oncol, 2011. **3**(2): p. 57-71.
2. Sprangers, M.A. and C.E. Schwartz, *Integrating response shift into health-related quality of life research: a theoretical model.* Soc Sci Med, 1999. **48**(11): p. 1507-15.
3. Dabakuyo, T.S., et al., *Response shift effects on measuring post-operative quality of life among breast cancer patients: a multicenter cohort study.* Qual Life Res, 2013. **22**(1): p. 1-11.
4. Andrykowski, M.A., K.A. Donovan, and P.B. Jacobsen, *Magnitude and correlates of response shift in fatigue ratings in women undergoing adjuvant therapy for breast cancer.* J Pain Symptom Manage, 2009. **37**(3): p. 341-51.
5. Schwartz, C.E. and M.A. Sprangers, *Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research.* Soc Sci Med, 1999. **48**(11): p. 1531-48.
6. Korfage, I.J., H.J. de Koning, and M.L. Essink-Bot, *Response shift due to diagnosis and primary treatment of localized prostate cancer: a then-test and a vignette study.* Qual Life Res, 2007. **16**(10): p. 1627-34.
7. Oort, F.J., *Using structural equation modeling to detect response shifts and true change.* Qual Life Res, 2005. **14**(3): p. 587-98.
8. Li, Y. and B. Rapkin, *Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS.* J Clin Epidemiol, 2009. **62**(11): p. 1138-47.
9. Boucekine, M., et al., *Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: a cohort study.* BMC Med Res Methodol, 2013. **13**: p. 20.
10. Oort, F.J., M.R. Visser, and M.A. Sprangers, *An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery.* Qual Life Res, 2005. **14**(3): p. 599-609.
11. King-Kallimanis, B.L., et al., *Structural equation modeling of health-related quality-of-life data illustrates the measurement and conceptual perspectives on response shift.* J Clin Epidemiol, 2009. **62**(11): p. 1157-64.

12. King-Kallimanis, B.L., et al., *Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients.* Qual Life Res, 2011. **20**(10): p. 1527-40.
13. Ware, J.E., Jr. and C.D. Sherbourne, *The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection.* Med Care, 1992. **30**(6): p. 473-83.
14. Teresi, J.A., *Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics.* Med Care, 2006. **44**(11 Suppl 3): p. S152-70.
15. Aaronson, N.K., et al., *The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology.* J Natl Cancer Inst, 1993. **85**(5): p. 365-76.
16. Sprangers, M.A., et al., *The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: first results from a three-country field study.* J Clin Oncol, 1996. **14**(10): p. 2756-68.
17. King-Kallimanis, B.L., et al., *Assessing measurement invariance of a health-related quality-of-life questionnaire in radiotherapy patients.* Qual Life Res, 2012. **21**(10): p. 1745-53.
18. Huang, C.C., et al., *Quality of life in Taiwanese breast cancer survivors with breast-conserving therapy.* J Formos Med Assoc, 2010. **109**(7): p. 493-502.
19. Fayers PM, et al., *EORTC QLQ-C30 Scoring Manual (3rd edition). Brussels: EORTC 2001 ed2001.* 2001.
20. Wilson, I.B. and P.D. Cleary, *Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes.* JAMA, 1995. **273**(1): p. 59-65.
21. Browne, M.W. and R. Cudeck, *Alternative ways of assessing model fit.* Sociological Methods & Research, 1992. **21**(2): p. 230-258.
22. Bentler, P.M., *Comparative fit indexes in structural models.* Psychological bulletin, 1990. **107**(2): p. 238.
23. De Ayala, R.J., *The theory and practice of item response theory. New York : Guilford Press,.* 2009.

24. Nolte, S., et al., *Tests of measurement invariance failed to support the application of the "then-test".* J Clin Epidemiol, 2009. **62**(11): p. 1173-80.
25. Hays, R.D., L.S. Morales, and S.P. Reise, *Item response theory and health outcomes measurement in the 21st century.* Med Care, 2000. **38**(9 Suppl): p. II28-42.
26. Fischer, G.H. and I.W. Molenaar, *Rasch models: Foundations, recent developments, and applications*. 1995: Springer.
27. Van der Linden, W.J. and R.K. Hambleton, *Handbook of Modern Item Response Theory*. Springer Verlag, New York ed. 1997.

Figures and Tables

Figure 1: Path diagram of the final measurement model integrating both QLQ-C30 and QLQ-BR23 dimensions and considering that the patient's symptomatic state (Sym) impacts his functional state (Fun).



Circles represent latent variable and residual factor and squares represent observed variables. Uni-item dimensions were excluded. HRQoL scores retained are Global Health Status (GHS), Physical Functioning (PF), Role Functioning (RF), Emotional Functioning (EF), Cognitive Functioning (CF), Social Functioning (SF), Body image (BI), Fatigue (FA), Nausea and Vomiting (NA), Pain (PA), Systemic Therapy Side Effects (STSE), Breast symptoms (BS) and Arm Symptoms (AS).

Table 1: Patients' baseline characteristics

	N	%
Hospital		
Dijon	271	71.1
Nancy	74	19.4
Reims	18	4.7
Strasbourg	18	4.7
Inclusion criteria		
confirmed primary breast cancer	242	63.5
suspicion of primary breast cancer	138	36.2
missing	1	0.3
Cancer		
confirmed	340	89.2
not confirmed	38	10.0
missing	3	0.8
Lymph node dissection (LND)		
Axillary LND	138	36.2
Sentinel lymph node biopsy	131	34.4
ALND+SLNB	32	8.4
No LND	75	19.7
missing	5	1.3
Surgery type		
mastectomy	124	32.6
no mastectomy	241	63.3
missing	16	4.2
Chemotherapy		
yes	155	40.7
no	218	57.2
missing	8	2.1
Radiotherapy		
yes	254	66.7
no	119	31.2
missing	8	2.1
Hormone therapy		
yes	170	44.6
no	203	53.3
missing	8	2.1

Table 2: Goodness of fit of models in the three step Response Shift detection procedure for the three analyses

Time points	Model	Description	Chi-square	DF	RMSEA	CFI
Baseline - after surgery	Model 1	With RS	1597.0	296	0.108	0.737
	Model 2	No RS	2455.5	332	0.130	0.571
	Model 3	Final model	1622.0	312	0.106	0.735
Baseline - 3 months	Model 1	With RS	1508.5	296	0.104	0.769
	Model 2	No RS	1999.0	332	0.115	0.682
	Model 3	Final model	1523.6	313	0.101	0.769
3 months - 6 months	Model 1	With RS	2070.6	296	0.131	0.723
	Model 2	No RS	2151.8	332	0.125	0.716
	Model 3	Final model	2099.2	325	0.125	0.723

RS: Response Shift

Table 3 : Parameters estimated in the final SEM model comparing measures at baseline and just after surgery (N=377 patients)

	Functional latent variable						Symptomatic latent variable						
	GHS	**PF**	**RF**	**EF**	**CF**	**SF**	**BI**	**FA**	**NA**	**PA**	**STSE**	**BS**	**AS**
Factor loading													
Baseline	1.2	1.1	1.6	1.0	1.1	1.4	0.9	17.9	6.2	16.9	10.5	5.4	8.6
After surgery	1.2	1.1	**2.2**	1.0	1.1	1.4	0.9	17.9	**3.8**	16.9	**7.3**	5.4	**11.2**
Intercept													
Baseline	69.9	90.2	89.6	64.9	83.5	90.9	89.2	22.6	3.4	12.4	12.9	11.0	7.7
After surgery	69.9	90.2	89.6	**80.5**	**90.6**	90.9	89.2	**19.5**	0.9	12.4	**8.2**	**15.5**	**10.0**
Residual variance													
Baseline	216.6	96.5	116.4	535.5	285.0	173.4	245.5	165.5	121.2	184.6	110.5	208.6	129.1
After surgery	216.6	**168.4**	**292.6**	**285.0**	285.0	**325.8**	**486.1**	**172.4**	121.2	184.6	110.5	208.6	**204.4**

HRQoL scores retained are Global Health Status (GHS), Physical Functioning (PF), Role Functioning (RF), Emotional Functioning (EF), Cognitive Functioning (CF), Social Functioning (SF), Body image (BI), Fatigue (FA), Nausea and Vomiting (NA), Pain (PA), Systemic Therapy Side Effects (STSE), Breast Symptoms (BS) and Arm Symptoms (AS).

Table 4: Parameters estimated in the final Structural Equation Model comparing measures at baseline and at three months post surgery (N = 378 patients)

	Functional latent variable						Symptomatic latent variable						
	GHS	**PF**	**RF**	**EF**	**CF**	**SF**	**BI**	**FA**	**NA**	**PA**	**STSF**	**BS**	**AS**
Factor loading													
Baseline	1.3	1.1	1.6	1.3	1.1	1.3	0.9	18.9	6.3	15.2	11.0	5.5	8.5
3 months	1.3	1.1	**2.0**	1.3	1.1	**1.8**	**1.3**	18.9	6.3	15.2	11.0	**8.4**	8.5
Intercept													
Baseline	68.8	90.7	89.7	64.8	83.4	90.7	89.6	22.8	4.2	12.3	13.6	11.1	8.08
3 months	**72.3**	90.7	**88.3**	**83.6**	**89.3**	90.7	**83.3**	22.8	4.2	12.3	13.6	**17.0**	8.08
Residual variance													
Baseline	199.1	95.8	121.7	528.2	304.2	176.2	245.9	153.5	130.3	183.6	118.9	188.2	127.9
3 months	199.1	95.8	**204.6**	**351.6**	304.2	**242.9**	**653.8**	153.5	**267.8**	**296.9**	**206.7**	**403.6**	**227.7**

HRQoL scores retained are Global Health Status (GHS), Physical Functioning (PF), Role Functioning (RF), Emotional Functioning (EF), Cognitive Functioning (CF), Social Functioning (SF), Body image (BI), Fatigue (FA), Nausea and Vomiting (NA), Pain (PA), Systemic Therapy Side Effects (STSE), Breast Symptoms (BS) and Arm Symptoms (AS).

Table 5: Parameters estimated in the final SEM model comparing measures at three months and six months post surgery (N=349 patients)

	Functional latent variable						Symptomatic latent variable					
	GHS	**PF**	**RF**	**EF**	**CF**	**SF**	**BI**	**FA**	**NA**	**PA**	**STSF**	**BS**	**AS**
Factor loading													
3 months	1.1	0.96	1.6	1.1	1.1	1.5	1.2	23.3	7.0	19.7	13.7	10.4	11.5
6 months	1.1	0.96	1.6	1.1	1.1	1.5	1.2	23.3	7.0	19.7	13.7	10.4	11.5
Intercept													
3 months	61.7	81.3	72.0	72.9	79.9	76.7	71.7	38.2	9.9	24.5	23.6	23.7	15.6
6 months	64.6	81.3	72.0	72.9	79.9	76.7	71.7	38.2	6.9	28.4	23.6	23.7	19.2
Residual variance													
3 months	167.6	95.7	195.1	337.7	309.7	240.6	630.5	169.1	268.4	290.9	191.5	405.9	221.4
6 months	167.6	95.7	195.1	337.7	309.7	**184.2**	630.5	169.1	**180.8**	290.9	191.5	**312.1**	221.4

HRQoL scores retained are Global Health Status (GHS), Physical Functioning (PF), Role Functioning (RF), Emotional Functioning (EF), Cognitive Functioning (CF), Social Functioning (SF), Body image (BI), Fatigue (FA), Nausea and Vomiting (NA), Pain (PA), Systemic Therapy Side Effects (STSE), Breast Symptoms (BS) and Arm Symptoms (AS).

V. DISCUSSION

En cancérologie, les résultats des analyses de données de QdV restent encore peu utilisés en pratique clinique pour changer les standards de prise en charge des patients. Ce manque de considération de la QdV est dû à la nature subjective des données de QdV et à la complexité de l'analyse longitudinale de la QdV. La QdV est un concept subjectif et dynamique, reflété par un effet « Response Shift », et potentiellement impacté par l'occurrence de données manquantes dépendantes des caractéristiques des patients et/ou de leur niveau de QdV manquant. De nombreuses méthodes statistiques ont été proposées mais peu de recommandations ont été données sur les approches statistiques à utiliser selon les situations thérapeutiques.

1. Le temps jusqu'à détérioration d'un score de QdV

Un premier objectif de ce travail a été d'investiguer la méthode du temps jusqu'à détérioration d'un score de QdV en tant que modalité d'analyse longitudinale de la QdV en cancérologie. Cette méthode est de plus en plus souvent utilisée dans les essais de phase III en cancérologie. Néanmoins, peu d'études ont été menées dans les essais de phase précoce et cette méthode n'avait jamais été appliquée à un essai de phase I.

Le TJD nécessite une définition de son évènement, soit la détérioration. Cette définition dépend du choix du score de référence, de la valeur de la DMCI retenue, de la prise compte ou non du décès dans la définition de l'évènement et peut correspondre à un état transitoire ou récurrent. La multiplicité des définitions possibles induit certaines recommandations sur les définitions à appliquer selon les situations thérapeutiques afin de pouvoir comparer les résultats entre les essais. Une utilisation massive de cette approche nécessite également une implémentation des différentes définitions possibles sous un logiciel de statistiques en libre accès.

Les données manquantes sont généralement ignorées dans cette approche. Le niveau de QdV est supposé constant entre deux données de QdV disponibles. Si une détérioration est suivie de données manquantes monotones, alors la détérioration peut être considérée comme une

détérioration définitive. L'impact des données manquantes sur l'analyse longitudinale de la QdV a déjà été démontré (Fairclough *et al*, 1998). Les données manquantes de type MAR peuvent être considérées comme des données manquantes de type MCAR conditionnellement aux covariables observées influençant l'occurrence des données manquantes. Il est important d'étudier l'impact de ces données manquantes sur la méthode du TJD et d'en tenir compte dans l'analyse.

Un autre enjeu de l'analyse longitudinale est l'occurrence potentielle de l'effet « Response Shift ». Bien que plusieurs méthodes statistiques existent pour mettre en évidence l'occurrence de ce phénomène, peu de recherches ont été menées sur la façon de prendre en compte l'occurrence d'un tel effet dans l'analyse longitudinale. Une alternative pour prendre en compte de façon indirecte l'occurrence d'un tel effet selon la méthode du TJD devait donc être proposée.

- **TJD et critères RECIST**

Compte tenu de la multiplicité des définitions possibles, un premier travail a été mené sur la proposition de critères RECIST (« Response Evaluation Criteria In Solid Tumors ») pour la QdV selon la méthode du TJD. Quelques recommandations ont été données sur les définitions de TJD à appliquer selon les situations thérapeutiques. Ainsi, en situation adjuvante, le TJD simple en tant qu'état transitoire semble la définition la plus adaptée puisque le patient a de grandes chances de guérir de son cancer. En revanche, en situation avancée ou métastatique, il parait plus pertinent d'étudier le TJD définitif, intégrant ou non le décès dans la définition de l'évènement, selon la définition de Bonnetain *et al.* (Bonnetain *et al*, 2010). De plus, certaines propositions pour tenir compte de façon indirecte de l'effet Response Shift ont été formulées : si un effet « Response Shift » est suspecté ou démontré, le score à l'inclusion pourrait ne pas être le score le plus adéquat pour qualifier une détérioration du niveau de QdV du patient. En revanche, le meilleur score antérieur ou le score immédiatement précédant pourrait être considéré comme le score de référence pour le patient.

Ces quelques recommandations sont une première étape dans la proposition d'une standardisation de l'analyse longitudinale selon la méthode du TJD. Nous devons désormais encourager l'utilisation de ces définitions et poursuivre les recherches sur une standardisation de la définition du TJD selon les situations thérapeutiques. De plus, des investigations

supplémentaires sur la capacité des scores de référence alternatifs tels que le meilleur score antérieur et le score immédiatement précédent pour tenir compte de l'effet « Response Shift » sont indispensables. Néanmoins, la possibilité de pouvoir choisir un score de référence autre que le score à l'inclusion est un atout indéniable de la méthode du TJD, contrairement à des méthodes plus classiques telles que les modèles mixtes.

Dans une optique de standardisation, un projet a été initié avec le groupe EORTC PROBE (« Patient Reported Outcome and Behavioural Evidence ») afin de proposer une standardisation de l'analyse de la QdV par le biais du développement de critères RECIST pour la QdV. Cette standardisation facilitera la comparaison entre les essais cliniques et permettra à long terme une meilleure considération des résultats de QdV.

- **TJD et package R**

Un second objectif a été de créer un package sous le logiciel R pour l'analyse longitudinale de la QdV selon la méthode du TJD. Ce package permet de réaliser une analyse du TJD d'un score de QdV selon un ensemble de définitions possibles en faisant varier le choix du score de référence (score à l'inclusion, meilleur score antérieur, score immédiatement précédent), le choix de la DMCI (en saisie directe par l'utilisateur) et la prise en compte ou non du décès dans la définition de l'évènement. La détérioration considérée peut être une détérioration simple en tant qu'état transitoire ou une détérioration définitive. Des analyses de sensibilités peuvent également être réalisées en parallèle de l'analyse principale. Ce package va permettre une large utilisation de l'approche du TJD. Ces analyses seront facilitées par l'implémentation des différentes définitions investiguées. L'aide créée pour ce package explique également à l'utilisateur le rôle de chaque paramètre des fonctions ainsi que les différentes valeurs possibles pour ces paramètres (Annexe C). Enfin, les exemples proposés pour chaque fonction facilitent la compréhension.

Ce package sera régulièrement complété et mis à jour. A titre d'illustration, il peut être pertinent, selon les situations, d'étudier la détérioration d'au moins un score de QdV parmi un ensemble de scores donnés. L'évènement correspond alors à l'occurrence d'une première détérioration, quel que soit le score concerné. L'implémentation de cette définition multi-scores sera réalisée dans une mise à jour du package et proposée en option des fonctions **TJD**

et **TUDD**. De plus, le scoring des modules de QdV récemment développés et validés de l'EORTC seront également implémentés et ajoutés aux fonctions de scoring déjà présentes dans le package.

- **TJD et macro SAS**

Certaines définitions du TJD ont également été programmées sous le logiciel de statistiques SAS sous la forme de deux macro-programmes. Ces programmes vont être complétés en implémentant l'ensemble des définitions de TJD investiguées et implémentées sous le logiciel R. Ils seront également mis à la disposition des utilisateurs du logiciel SAS. Ainsi, l'implémentation des définitions du TJD sous les logiciels SAS permettra de faciliter son analyse par les utilisateurs qui ne sont pas familiers avec le logiciel R.

- **TJD et données manquantes**

Un travail a également été mené sur la prise en compte des données manquantes de type MAR selon la méthode du TJD. L'utilisation de la méthode « inverse probability of treatment weighting » du score de propension a été investiguée, conjointement avec la méthode du TJD. Cette méthode a été appliquée à un essai de phase II portant sur le cancer du pancréas en situation métastatique. Cette méthode a l'avantage d'être facilement compréhensible et réalisable : il s'agit d'une pondération des observations selon la présence ou non de données manquantes au cours du suivi du patient. D'autres méthodes existent pour tenir compte des données manquantes de type MAR. Les imputations multiples tenant compte des facteurs influençant l'occurrence des données manquantes ont souvent été proposées (Fairclough, 2010). Cependant, les algorithmes utilisés pour réaliser ces imputations peuvent être difficiles à appréhender pour certains cliniciens. La réalisation de ces imputations peut également poser certaines difficultés méthodologiques mais aussi philosophiques : doit-on imputer uniquement les items manquants à hauteur de 50% ou bien également les scores manquants ? Comment tenir compte de la corrélation entre les données de QdV si l'imputation se fait sur les items par dimension ? Doit-on imputer les scores même après la sortie d'étude du patient ? Doit-on imputer des données après le décès du patient ? En outre, cette méthode peut rencontrer des difficultés de convergence si le nombre de patients n'est pas suffisant. L'essai que nous avons

considéré est un essai de phase II avec un nombre réduit de patients : la technique des imputations multiples semble donc peu adaptée pour cette étude. Enfin, cette méthode ne peut tenir compte que d'un nombre restreint de facteurs associés aux données manquantes en raison de ses difficultés de convergence. Tous ces éléments amènent à croire que la méthode « inverse probability of treatment weigthing » semble préférable à la méthode d'imputations multiples pour le TJD et particulièrement dans les essais à faible effectifs.

Afin de garantir la capacité de cette méthode à tenir compte de façon adéquate de l'occurrence des données manquantes de type MAR, nous allons désormais tester cette approche par le biais d'une étude de simulations. Ce projet est une collaboration avec une équipe italienne (« Italian Group for Adult Hematologic Diseases » (GIMEMA)). Pour ce faire, différents scénarios vont être explorés en se basant à la fois sur nos travaux déjà réalisés pour simuler les données de QdV longitudinales et sur les travaux réalisés par l'équipe italienne pour simuler des données manquantes de type MAR (Cottone *et al*, 2013). Les méthodes explorées pour tenir compte des données manquantes seront la méthode IPTW du score de propension, mais également la méthode d'appariement du score de propension et la méthode d'imputation multiple tenant compte des facteurs influençant la présence des données manquantes. Ces méthodes seront appliquées à la méthode du TJD en considérant deux définitions : le TJD simple (Hamidou *et al*, 2011) et le TJD définitif (Bonnetain *et al*, 2010). Les simulations seront réalisées avec un effet temps et soit sans effet bras de traitement (pour pouvoir déterminer le taux d'erreur de type I du test d'un effet bras de traitement) soit avec un effet bras de traitement (pour déterminer la puissance statistique). Les critères de comparaison seront :

- la puissance statistique et le taux d'erreur de type I du test d'un effet bras de traitement selon le test du Log-Rank ;
- la médiane de détérioration avec intervalle de confiance à 95% ;
- le coefficient Hazard Ratio entre les deux bras de traitement avec son intervalle de confiance à 95% ;
- La concordance des résultats avec la moyenne du coefficient Kappa de Cohen avec son intervalle de confiance à 95%, pour mesurer la cohérence entre les censures et évènements entre chaque méthode pour tenir compte des données manquantes et les résultats obtenus sur données complètes ;

- Et la moyenne du coefficient de corrélation de Spearman avec son intervalle de confiance à 95% pour comparer les temps de survies.

- **TJD et risques compétitifs**

Le TJD présente encore certaines difficultés méthodologiques. Des investigations supplémentaires vont être réalisées pour tester la robustesse de cette méthode d'analyse longitudinale, voire améliorer son estimation. Le TJD étant une méthode de type analyse de survie, elle peut être sujette à la présence de risques compétitifs qui peuvent impacter l'analyse s'ils ne sont pas correctement pris en compte. Ainsi, un travail en cours consiste à tester la présence de risques compétitifs entre les différents scores de QdV et le décès. Ces analyses sont réalisées sur plusieurs bases de données de phase II ou phase III, correspondant à diverses situations thérapeutiques et localisations cancéreuses. En présence de risques compétitifs, l'analyse doit être adaptée en conséquence. Par exemple, un test de Gray sera appliqué au lieu du test du Log-Rank pour comparer les résultats entre les bras de traitement (Gray, 1988); un modèle de Fine and Gray sera également appliqué à la place des modèles de Cox plus usuels (Fine & Gray, 1999).

- **TJD et essais de phase I**

La QdV est à ce jour particulièrement étudiée dans les essais de phase III. Elle commence également à être mesurée et étudiée dans les essais de phase II (Garsa *et al*, 2013; Gontero *et al*, 2013; Mustea *et al*, 2013). En revanche, elle reste encore peu exploitée dans les essais de phase I (Rouanne *et al*, 2013; Stephenson *et al*, 2013). Un des objectifs de ces essais est de déterminer la dose recommandée pour les essais de phase II. Ces essais sont réalisés sur un faible nombre de patients. La dose recommandée correspond à la dose la plus élevée que l'on puisse administrer aux patients tout en entrainant un niveau de toxicités acceptable pour le patient. Ces toxicités et leur intensité sont habituellement mesurées par les cliniciens selon la grille NCI-CTC AE (NCI, 2006). L'avis et la perception du patient pourraient compléter cette évaluation. La valeur ajoutée de la QdV dans les essais de phase I pour déterminer la dose maximale recommandée semble donc légitime. Dans cette optique, une étude exploratoire du TJD de la QdV a été réalisée dans un essai de phase I, tout en tenant compte du palier de dose

et de l'occurrence ou non de toxicités. Différentes définitions de détérioration ont été explorées, en considérant la détérioration d'un score ou la détérioration d'au moins un score de QdV, intégrant ou non l'occurrence d'une toxicité de grade3/4 comme évènement. Les résultats ont montré une cohérence entre le palier de dose retenu comme dose maximale tolérée et le TJD de la QdV, et ce quelle que soit la définition de la détérioration considérée : les patients du palier de dose retenue présentent une détérioration plus tardive de la QdV comparativement aux patients inclus dans les autres paliers de dose. La méthode du TJD semble appropriée à ce type d'essai puisqu'elle permet de tenir compte de l'occurrence des toxicités selon l'évaluation effectuée par le médecin. Ces résultats permettent également de compléter les résultats de l'étude principale concernant la détermination de la dose maximale tolérée. Les résultats sont ainsi renforcés par l'étude de la QdV selon le palier de dose. Cette étude soulève par ailleurs la nécessité de créer un questionnaire de QdV spécifique des essais précoces et qui serait davantage centré sur les symptômes.

2. Comparaison de trois approches statistiques pour l'analyse longitudinale

La méthode du TJD est de plus en plus utilisée en tant que modalité d'analyse longitudinale de la QdV dans les essais cliniques en cancérologie (Bonnetain *et al*, 2010; Burris *et al*, 2013; Kabbinavar *et al*, 2008). Elle présente certains avantages indéniables tels que la possibilité d'utiliser différents scores de référence en présence d'un effet « Response Shift » et son aptitude à s'adapter à la situation thérapeutique et à présenter des résultats familiers pour les cliniciens. Cependant, nous ne connaissons pas la capacité théorique de cette méthode à mettre en évidence une différence entre deux bras de traitements. Il était donc essentiel de comparer cette approche à une approche plus classique tel que le modèle linéaire à effets mixtes basé sur le score (Fairclough, 2010) sur des données de simulations. Ce modèle est en effet souvent utilisé pour l'analyse longitudinale des données de QdV (Penar-Zadarko *et al*, 2013; Stockler *et al*, 2014). Les modèles IRT sont également des modèles adaptés aux données issues des questionnaires de QdV et bien qu'encore peu exploités pour l'analyse longitudinale, ils seraient a priori plus adaptés pour l'analyse longitudinale de la QdV que les modèles mixtes (de Bock *et al*, 2013).

Un objectif de ce travail a donc été de comparer l'approche du TJD avec l'approche plus classique du modèle linéaire à effets mixtes et un modèle IRT longitudinal adapté aux items polytomiques. Cette comparaison s'est faite sur des données de simulations en investiguant la capacité de ces méthodes à mettre en évidence un effet d'interaction entre le temps et le traitement. Différents scénarios de simulations ont été explorés, proches du design des essais cliniques et de la construction des questionnaires de QdV de l'EORTC. Ainsi, les simulations ont été réalisées avec 5 ou 10 temps de mesure et en considérant que la QdV est évaluée par 1, 2 ou 4 items à 4 ou 7 modalités de réponse. De plus, afin de refléter les conditions réelles des essais cliniques, l'impact des données manquantes dépendant du niveau de QdV des patients (type MNAR) a également été étudié, en générant à la fois des données manquantes intermittentes et monotones de type MNAR.

Les simulations ont montré une performance équivalente entre le modèle linéaire mixte et le modèle IRT longitudinal sur les données complètes, avec une puissance plus faible pour les modèles IRT pour les scénarios avec 10 temps de mesure. En revanche, la méthode du TJD était très peu puissante pour les simulations réalisées avec un seul item, quelle que soit la définition de la détérioration considérée. Ces résultats suggèrent que la méthode du TJD devrait être utilisée avec prudence sur les dimensions de QdV évaluée par un seul item telle que la majorité des dimensions symptomatiques du questionnaire EORTC QLQ-C30. Cette méthode devrait à défaut être utilisée en complément d'une analyse plus performante telle que le modèle linéaire mixte ou le modèle IRT longitudinal.

En présence de données manquantes informatives, les simulations ont montré que les puissances statistiques de toutes les méthodes diminuent excepté celles du modèle IRT longitudinal qui auraient plutôt tendance à augmenter. Les modèles IRT seraient donc plus adaptés pour les études à faible temps de mesure ou du moins en présence de données manquantes. La puissance du TJD par rapport au meilleur score antérieur avait également tendance à augmenter en présence de données manquantes. Cependant, cette méthode reste très peu puissante. Néanmoins, il est important de noter que le décès, habituellement utilisé dans les définitions de TJD, n'a pas été simulé dans cette étude. La faible puissance du TJD peut donc en partie être expliquée par la non prise en compte de la variable décès.

Cette étude de simulation était la première étude de simulation menée :

- sur des items polytomiques ;
- avec plus de 3 temps de mesure ;
- testant un effet d'interaction entre le temps et l'effet traitement ;
- et étudiant l'occurrence conjointe de données manquantes intermittentes et monotones.

Ces scénarios ont l'avantage d'être proche du design des essais cliniques et des conditions réelles des essais. De plus, ils se basent sur la structure des questionnaires de l'EORTC souvent utilisés dans les essais cliniques en cancérologie.

Enfin, cette étude de simulation donne quelques recommandations sur les méthodes statistiques à utiliser pour l'analyse longitudinale de la QdV selon les données recueillies (nombre de temps de mesure, nombre d'items).

- **Etude impact Response Shift sur les trois méthodes pour l'analyse longitudinale**

Un autre enjeu de l'analyse longitudinale est l'occurrence potentielle d'un effet « Response Shift ». L'occurrence d'un tel effet peut biaiser l'analyse longitudinale. Ainsi, il parait également pertinent d'étudier l'impact de l'occurrence d'un effet « Response Shift » sur ces différentes approches par le biais de simulations.

Un projet a donc été initié sur l'étude de l'impact de la « Response Shift » sur ces méthodes statistiques par le biais de simulation. Ce projet se focalise sur la simulation de la composante recalibration de l'effet « Response Shift ». Cette composante de la Response Shift est la plus importante à prendre en compte dans l'analyse longitudinale puisqu'elle caractérise une différence au niveau de l'échelle de mesure. De plus, elle ne nécessite pas l'utilisation d'un modèle multidimensionnel, contrairement aux composantes changement de valeurs et de conceptualisation de la QdV. Enfin, étudier la composante « changements de valeurs » nécessiterait d'utiliser un modèle IRT dont le paramètre de discrimination puisse varier entre les items et les temps de mesure. Un tel modèle ne correspondait donc pas à un modèle de la famille des modèles de Rasch.

Différents scénarios de simulations sont élaborés, en considérant que l'effet recalibration affecte un seul des deux bras de traitement ou bien les deux bras de traitement. Cet effet recalibration correspond à un changement au niveau des paramètres de difficulté des items.

Ainsi, une augmentation du coefficient de difficulté de l'item pour une échelle symptomatique correspond à une recalibration à la baisse, i.e. que les patients avaient surestimé la présence de ce symptômes lors de la (ou des) mesure(s) antérieure(s).

Pour des scénarios considérant 5 temps de mesure, un effet recalibration pourra être simulé à partir du 3ème temps de mesure. Le patient étant en cours de traitement pour un cancer diagnostiqué à l'inclusion (première mesure), il pourra alors revoir son échelle d'évaluation de ses symptômes à la baisse, ou inversement son niveau de QdV à la hausse. La comparaison des différentes méthodes d'analyses se fera à nouveau selon l'erreur de type I et la puissance statistiques du test d'une interaction entre le temps et le bras de traitement.

- **Comparaison des trois méthodes statistiques sur données réelles**

Afin de compléter les résultats de ces études de simulations, nous allons également comparer les différentes méthodes statistiques pour l'analyse longitudinale sur données réelles issues de plusieurs essais cliniques correspondant à différentes situations thérapeutiques et localisations cancéreuses. Cette comparaison se fera alors d'un point de vue clinique en faisant appel à un panel d'experts pour leur demander leur avis quant à la performance des différentes méthodes et leur capacité à produire des résultats cliniquement pertinents.

- **Investigation de modèles non paramétriques pour l'analyse longitudinale**

Le modèle linéaire à effets mixtes est la méthode la plus utilisée pour l'analyse longitudinale des données de QdV. La performance de ce modèle a par ailleurs été démontrée par le biais de notre étude de simulations. Cependant, cette méthode repose sur l'hypothèse de normalité de la variable d'intérêt, soit du score de QdV. Cette hypothèse est rarement vérifiée dans les études de QdV. De plus, elle semble peu réaliste, en particulier pour les échelles uni-item respectant davantage une propriété d'ordonnancement. Ainsi, il parait primordial d'investiguer des modèles non paramétriques pour l'analyse longitudinale de la QdV. Peu de recherches sont menées à ce jour sur la modélisation de données de QdV longitudinales par un modèle non paramétrique tel qu'un modèle non paramétrique par noyau (Kokonendji & Kiessé, 2011). Ainsi, un projet va être mené sur ces modèles et leur capacité à modéliser un

effet temps dans le cadre de données ne respectant pas une distribution normale. Ces modèles étant complexes, il sera indispensable de proposer des résultats accessibles aux cliniciens afin de faciliter la prise en compte des résultats en pratique clinique.

3. Caractérisation de l'occurrence de l'effet Response Shift

Un second objectif de mon travail a été d'investiguer différentes approches statistiques pour caractériser l'occurrence de l'effet « Response Shift ». Un premier travail a consisté à explorer l'utilisation des analyses factorielles et modèles IRT pour caractériser l'occurrence de la « Response Shift ». Ainsi, la composante recalibration de l'effet « Response Shift » a été explorée par des Analyses Factorielles des Correspondances Multiples (AFCM) et des modèles IRT conjointement avec la méthode Then-test. Les composantes changements de valeurs et de conceptualisation de la QdV ont été explorées par des Analyses en Composante Principales (ACP). Les AFCM et les modèles IRT ont montré des résultats cohérents et complémentaires pour caractériser la recalibration. Cependant, une limite de ces méthodes est qu'elles ont été appliquées conjointement avec la méthode Then-test. Bien que nous ayons appliqué les recommandations données par Schwartz *et al.* pour améliorer les analyses effectuées selon la méthode Then-test (Schwartz & Sprangers, 2010), cette méthode présente certains inconvénients comme un possible biais de mémoire et de désirabilité sociale (McPhail & Haines, 2010; Nolte *et al*, 2009). De plus, cette méthode nécessite de prévoir une évaluation de l'effet « Response Shift » dès le design de l'étude. Les ACP ont permis de mettre en évidence les composantes changements de valeurs et de conceptualisation de l'effet « Response Shift » en se basant uniquement sur les questionnaires prospectifs. Cette méthode a l'avantage d'offrir une visualisation graphique de ces deux composantes et des liaisons entre les différents scores au cours du temps. Cependant, les ACP ne sont pas des modèles de mesure et ne tiennent pas compte de la corrélation entre les données au cours du temps.

Les modèles à équations structurelles (SEM) sont une généralisation des ACP et correspondent à un modèle de mesure. En 2005, Oort a proposé l'utilisation de ces modèles pour caractériser les trois composantes de l'effet « Response Shift » (Oort, 2005) et a ensuite proposé une application de ces modèles dans une étude où la QdV a été évaluée par le questionnaire générique SF-36 (Oort *et al*, 2005). Par la suite, la procédure de Oort a été

appliquée dans de nombreuses études pour caractériser l'occurrence de l'effet « Response Shift » (King-Kallimanis *et al*, 2011; King-Kallimanis *et al*, 2009). Cependant, toutes ces études ont été réalisées sur le questionnaire générique SF-36. En cancérologie, la QdV est généralement mesurée par le biais de questionnaires spécifiques du cancer et principalement par les questionnaires EORTC QLQ-C30 ou FACT-G. Ainsi, un second travail sur l'effet « Response Shit » a été mené en investiguant les modèles SEM pour caractériser l'occurrence de l'effet « Response Shift » lorsque la QdV a été évaluée par les questionnaires de l'EORTC. Ces modèles SEM ont l'avantage de se focaliser sur les mesures prospectives et ainsi de ne pas dépendre de la mesure Then-test. Dans cette étude, ces modèles ont permis de mettre en évidence les composantes recalibration et changements de valeurs de la Response Shift. La composante reconceptualisation de la QdV n'a pas pu être explorée compte tenu de la structure des questionnaires de l'EORTC et de la construction de notre modèle de mesure. En effet, une investigation de la reconceptualisation nécessite une décomposition de la QdV en au moins deux domaines clairement distincts, ce qui n'est pas le cas des questionnaires EORTC. En revanche, le questionnaire SF-36 respecte une telle structure avec certaines dimensions de QdV évaluant un domaine de santé mentale et les autres dimensions évaluant un domaine de santé physique.

Les modèles SEM présentent des résultats assez cohérents avec ceux obtenus par les modèles IRT via la méthode then-test pour ce qui est de la recalibration uniforme. A titre d'illustration, pour l'analyse effectuée après la première hospitalisation, les modèles SEM et les modèles IRT mettent en évidence une sous-estimation des scores de fonctionnent émotionnel et cognitif et des symptômes au niveau du bras à l'inclusion, ainsi qu'une surestimation du score de nausée et vomissement. En revanche, à 6 mois comparativement à 3 mois, peu de recalibration est mise en évidence par les modèles SEM tandis que les IRT mettent en évidence une recalibration à la baisse pour la majorité des scores de QdV. Cette différence de résultats peut s'expliquer par le mauvais ajustement du modèle SEM final.

Les AFCM ont également permis de mettre en évidence des profils de recalibration à la baisse ou à la hausse, généralement d'une modalité de réponse à une modalité de réponse adjacente. Ces analyses pourraient permettre de refléter la composante recalibration non uniforme selon la terminologie des modèles SEM. Cependant, ces résultats peuvent difficilement être comparés à ceux obtenus par les modèles SEM puisqu'ils offrent davantage une

représentation graphique des profils de recalibration plutôt qu'une mesure quantitative de cette composante.

Concernant la composante « changement de valeurs », l'augmentation de l'importance relative des échelles fonctionnelles démontrées par les ACP est retrouvée pour les échelles de fonctionnement de rôle et fonctionnement social pour les SEM. Les deux analyses (ACP et SEM) mettent également en évidence une augmentation de l'importance relative de la dimension des symptômes au niveau du bras. L'absence de « changements de valeurs » à 6 mois comparativement à 3 mois de la QdV a également été mise en évidence par les deux types d'analyses.

Il est important de noter que ces différentes analyses n'ont pas pu être menées sur les mêmes échantillons de patientes. Ainsi, les résultats sont difficilement comparables. En effet, dans les modèles IRT et les AFCM, seules les patientes présentant une DMCI d'au moins 5 points entre les mesures pré-test et then-test étaient retenues et l'analyse a été menée score par score. Pour les ACP, les patientes ayant l'ensemble des données disponibles aux 4 temps de mesures ont été retenues. Enfin, les modèles SEM ont été appliqués sur l'ensemble des patients pour lesquelles au moins un score était disponible pour l'une des deux mesures pré-test ou post-test.

De plus, les échelles uni-items ont dû être exclues des analyses selon le modèles SEM alors qu'elles ont pu être exploitées dans notre précédente étude par analyses factorielles et modèles IRT. Ces analyses ont par ailleurs montré une recalibration importante pour certaines de ces échelles et également un changement de valeurs pour les dimensions diarrhée et perspectives futures.

Les modèles SEM semblent peu adaptés aux questionnaires EORTC dont la majorité des dimensions symptomatiques sont évaluées par un seul item. Ces dimensions ne respectent donc pas de distribution normale du score. Or, l'estimation des modèles sur des échantillons de cette taille (381 patientes incluses dans l'étude) se fait nécessairement par la méthode de maximum de vraisemblance en supposant la normalité des variables. Ces modèles seraient ainsi plus appropriés au questionnaire SF-36. Ils pourraient également être adaptés aux questionnaires de QdV du groupe FACT. En effet, ces questionnaires évaluent 4 à 5

dimensions de QdV et chaque dimension est évaluée par au moins 6 items à 5 modalités de réponse chacun. Un score global correspondant à la somme des scores obtenus pour chaque dimension est alors calculé, correspondant à un niveau de QdV global.

Pour ce qui est des questionnaires de l'EORTC, il parait plus pertinent et adapté de continuer à investiguer la capacité des modèles IRT pour caractériser l'occurrence de l'effet « Response Shift ». En effet, ces modèles sont beaucoup plus adaptés aux échelles de QdV, et particulièrement aux échelles uni-items. De ce fait, un travail va être mené sur la capacité des modèles IRT à mettre en évidence l'occurrence de l'effet Response Shift sans la nécessité de la mesure Then-test. Un modèle IRT de la famille des modèles de Rasch semble peu envisageable puisque ces modèles imposent un coefficient de discrimination constant égal à 1 pour tous les items. Or, la composante changement de valeurs du Response Shift caractérise un changement de l'importance relative des différentes dimensions de QdV. Un coefficient de discrimination constant impose que chaque item ait la même importance au regard du trait latent, soit du niveau de QdV du patient, ce qui ne permettrait pas de caractériser cette composante. Ainsi, les modèles de la famille de Lord seraient a priori plus appropriés pour ce type d'analyses. Les modèles « Generalized Partial Credit Model » (GPCM) et « Graded Response Model » (GRM) sont deux modèles de la famille de Lord relâchant l'hypothèse d'un coefficient de discrimination constant entre les items et adaptés aux items polytomiques. Ces modèles pourraient donc être explorés pour caractériser l'occurrence de l'effet Response Shift dans une étude évaluant la QdV par le biais des questionnaires de l'EORTC. De plus, les composantes changements de valeurs et de conceptualisation de la QdV font appel au caractère multidimensionnel de la QdV. Une utilisation des modèles IRT multidimensionnels pour caractériser l'effet « Response Shift » est donc requise. Peu de modèles IRT multidimensionnels ont été développés à ce jour (Fisher, 1995; Hartig & Höhler, 2009; Reckase, 2009). Le modèle LLRA que j'ai appliqué pour caractériser la recalibration conjointement avec la méthode Then-test est un modèle multidimensionnel. Néanmoins, l'estimation de ce modèle se fait par la méthode d'estimation de maximum de vraisemblance conditionnelle qui ne tient pas compte de l'occurrence des données manquantes. D'autres modèles multidimensionnels, dont l'estimation peut se faire par la méthode de maximum de vraisemblance marginale par exemple, doivent donc être investigués.

Dans ce futur projet, différents modèles IRT vont donc être étudiés : les modèles GPCM et GRM ainsi que les modèles IRT multidimensionnels adaptés aux items polytomiques. Ces

modèles seront explorés sur les questionnaires EORTC dans la même étude que celle dans laquelle nous avons investigué les analyses factorielles, les IRT et modèles SEM.

Une comparaison sera ensuite réalisée entre les approches IRT d'une part et entre ces modèles et les autres méthodes investiguées précédemment (analyses factorielles et IRT via le then-test, ACP, SEM).

Pour finir, les premières recherches que nous avons effectuées étaient dépendantes de la méthode Then-test. Cette méthode a déjà montré par le passé certains inconvénients tels qu'un possible biais de mémoire ou de désirabilité sociale (McPhail & Haines, 2010). Ainsi, cette mesure ne serait pas fiable pour mesurer l'effet Response Shift seul. Nolte *et al*. ont proposé de tester l'invariance de la mesure Then-test via l'utilisation des modèles SEM et en adaptant la procédure de Oort (Nolte *et al*, 2009). La QdV ayant été mesurée par des questionnaires de QdV de l'EORTC dans notre étude, nous proposons d'utiliser les modèles IRT pour tester l'invariance de la mesure Then-test. Ainsi, les mêmes modèles que ceux que nous allons investiguer pour caractériser l'occurrence de la Response Shift pourront être également utilisés pour tester l'invariance de la mesure Then-test. Cette analyse permettra en outre de conforter les résultats de nos analyses menées conjointement avec la méthode Then-test pour caractériser la composante recalibration de l'effet « Response Shift ».

VI. CONCLUSIONS

Ce travail a porté sur l'analyse longitudinale de la QdV en cancérologie avec en particulier deux grandes thématiques abordées: la caractérisation de l'occurrence de l'effet « Response Shit » et la méthode du temps jusqu'à détérioration en tant que modalité d'analyse longitudinale de la QdV.

Les différents travaux menés permettent de donner quelques premières recommandations, tant au niveau des méthodes statistiques pour caractériser l'occurrence de l'effet « Response Shift » que pour la méthodologie d'analyse longitudinale de la QdV.

Ainsi, les modèles à équations structurelles, souvent utilisés pour caractériser l'occurrence de l'effet « Response Shift », ne seraient pas une méthode adaptée à la structure des questionnaires de QdV de l'EORTC. Cette méthode ne doit donc pas être considérée comme un standard et d'autres méthodes doivent être proposées telles que les modèles IRT lorsque la QdV est évaluée par un questionnaire de l'EORTC.

Bien que différentes méthodes pour caractériser l'effet « Response Shift » existent, peu de recherches ont été menées sur la façon dont on doit tenir compte de l'occurrence de cet effet dans l'analyse longitudinale de la QdV. Dans le cadre d'une analyse longitudinale réalisée selon la méthode du TJD, l'utilisation d'un score de référence dynamique a été proposée pour qualifier la détérioration en présence d'un effet « Reponse Shift » avec le meilleur score antérieur ou le score immédiatement précédent comme score de référence. Des investigations supplémentaires sont à effectuer afin de garantir la capacité de ces scores alternatifs à tenir compte indirectement de l'occurrence d'un tel effet.

Le TJD est de plus en plus souvent utilisé en tant que méthode d'analyse longitudinale de la QdV dans les essais de phase III en cancérologie. Cette méthode à l'avantage, contrairement à d'autres méthodes d'analyses longitudinales, de pouvoir adapter la définition de l'évènement selon la situation thérapeutique, en considérant soit une détérioration simple en tant qu'état transitoire en situation adjuvante, soit une détérioration définitive en situation avancée ou métastatique. D'autre part, elle offre des résultats qui sont familiers pour les cliniciens, résumés sous forme d'Hazard Ratio et de médianes de survie. La création du package R pour l'analyse du TJD permet de facilité la réalisation de ces analyses et va également permettre de contribuer à une large diffusion de cette approche. Cette méthode a été investiguée à titre

exploratoire dans le cadre d'un essai de phase I et à montrer des résultats cohérents entre le palier de dose retenue comme dose maximale tolérée selon l'évaluation de la toxicité effectuée par le médecin et l'analyse de la QdV ressentie par le patient. Une méthode statistique a également été proposée pour tenir compte de l'occurrence des données manquantes de type MAR conjointement avec la méthode du TJD. Cette méthode a été appliquée dans le cadre d'un essai de phase II et devra être validée par le biais d'une étude de simulation.

Par ailleurs, l'étude de simulations que j'ai réalisée dans ce travail a permis de mettre en évidence le manque de puissance de cette approche pour les échelles de QdV mesurer par un seul item comme c'est le cas pour certaines dimensions symptomatiques du questionnaire EORTC QLQ-C30 et de ces modules supplémentaires. Il est donc nécessaire d'analyser ce type d'échelle selon un modèle linéaire mixte ou un modèle d'IRT longitudinal.

En conclusion, ce travail a permis de donner quelques recommandations pour l'analyse de la QdV en cancérologie, tant au niveau de la caractérisation de l'effet « Response Shift » que sur la méthodologie d'analyse longitudinale. Cependant, des investigations supplémentaires sont à effectuer afin de garantir une analyse longitudinale rigoureuse et robuste, à la fois aux données manquantes mais aussi à la présence d'un effet « Response Shift ». Au final, ces recherches permettront de standardiser l'analyse longitudinale de la QdV dans les essais cliniques selon les situations thérapeutiques et le design des études, ce qui facilitera les comparaisons entre les différents essais. Cette standardisation permettra une meilleure prise en compte des résultats de QdV en pratique clinique. Enfin, elle rendra également plus abordable la prise en compte de la QdV en tant que co-critère de jugement principal avec un critère tumoral dans certaines situations thérapeutiques. Pour cela, il reste néanmoins à mettre en œuvre une méthode statistique rigoureuse pour le calcul du nombre de sujets nécessaires dans le cas de critères de jugements multiples intégrant des données de QdV et de survie.

VII. REFERENCES

Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JC, et al. (1993) The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* **85**(5)**:** 365-76

Addington-Hall J, Kalra L (2001) Who should measure quality of life? *BMJ* **322**(7299)**:** 1417-20

Ahmed S, Mayo NE, Corbiere M, Wood-Dauphinee S, Hanley J, Cohen R (2005) Change in quality of life of people with stroke over time: true change or response shift? *Qual Life Res* **14**(3)**:** 611-27

Ahmed S, Sawatzky R, Levesque JF, Ehrmann-Feldman D, Schwartz CE (2014) Minimal evidence of response shift in the absence of a catalyst. *Qual Life Res*

Amir E, Seruga B, Kwong R, Tannock IF, Ocana A (2012) Poor correlation between progression-free and overall survival in modern clinical trials: are composite endpoints the answer? *Eur J Cancer* **48**(3)**:** 385-8

Andrich D (1978) A rating formulation for ordered response categories. *Psychometrika* **43**(4)**:** 561-573

Andrykowski MA, Donovan KA, Jacobsen PB (2009) Magnitude and correlates of response shift in fatigue ratings in women undergoing adjuvant therapy for breast cancer. *J Pain Symptom Manage* **37**(3)**:** 341-51

Austin PC (2013) The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med* **32**(16)**:** 2837-49

Awad L, Zuber E, Mesbah M (2002) Applying survival data methodology to analyze longitudinal quality of life data. In *Statistical Methods for Quality of Life Studies*, pp 231-243. Springer

Bacci S (2008) Analysis of longitudinal HrQoL using latent regression in the context of Rasch modeling. *Mathematical Methods in Survival Analysis, Reliability and Quality of Life***:** 275-290

Barclay-Goddard R, Epstein JD, Mayo NE (2009) Response shift: a brief overview and proposed research priorities. *Qual Life Res* **18**(3)**:** 335-46

Beauchemin C, Cooper D, Lapierre ME, Yelle L, Lachaine J (2014) Progression-free survival as a potential surrogate for overall survival in metastatic breast cancer. *Onco Targets Ther* **7:** 1101-10

Beitz J, Gnecco C, Justice R (1996) Quality-of-life end points in cancer clinical trials: the U.S. Food and Drug Administration perspective. *J Natl Cancer Inst Monogr*(20)**:** 7-9

Bellera CA, Pulido M, Gourgou S, Collette L, Doussau A, Kramar A, Dabakuyo TS, Ouali M, Auperin A, Filleron T, Fortpied C, Le Tourneau C, Paoletti X, Mauer M, Mathoulin-Pelissier S, Bonnetain F (2013) Protocol of the Definition for the Assessment of Time-to-event Endpoints in CANcer trials (DATECAN) project: formal consensus method for the development of guidelines for standardised time-to-event endpoints' definitions in cancer clinical trials. *Eur J Cancer* **49**(4)**:** 769-81

Bergman B, Aaronson NK, Ahmedzai S, Kaasa S, Sullivan M (1994) The EORTC QLQ-LC13: a modular supplement to the EORTC Core Quality of Life Questionnaire (QLQ-C30) for use in lung cancer clinical trials. EORTC Study Group on Quality of Life. *Eur J Cancer* **30A**(5)**:** 635-42

Blanchin M, Hardouin J-B, Le Neel T, Kubis G, Sebille V (2011a) Analysis of longitudinal Patient-Reported Outcomes with informative and non-informative dropout: Comparison of CTT and Rasch-based methods. *International Journal of Applied Mathematics & Statistics [Internet]* **24:** I-11

Blanchin M, Hardouin JB, Guillemin F, Falissard B, Sebille V (2013) Power and sample size determination for the group comparison of patient-reported outcomes with Rasch family models. *PLoS One* **8**(2)**:** e57279

Blanchin M, Hardouin JB, Le Neel T, Kubis G, Blanchard C, Mirallie E, Sebille V (2011b) Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes. *Stat Med* **30**(8)**:** 825-38

Blazeby J, Sprangers MA, Cull A, Grønvold M, Bottomley A (2001) *EORTC Quality of Life Group Guidelines for Developing Questionnaire Modules, 3'rd Edition*: EORTC. Report no. 2930064242

Boisson V. Test de type-log rank pour l'évolution de la qualité de vie liée à la santé. Thèse Université Pierre et Marie Curie-Paris VI, 2008

Bonnetain F (2010) Health related quality of life and endpoints in oncology. *Cancer Radiother* **14**(6-7)**:** 515-518

Bonnetain F, Bosset JF, Gerard JP, Calais G, Conroy T, Mineur L, Bouche O, Maingon P, Chapet O, Radosevic-Jelic L, Methy N, Collette L (2012) What is the clinical benefit of preoperative chemoradiotherapy with 5FU/leucovorin for T3-4 rectal cancer in a pooled analysis of EORTC 22921 and FFCD 9203 trials: surrogacy in question? *Eur J Cancer* **48**(12)**:** 1781-90

Bonnetain F, Dahan L, Maillard E, Ychou M, Mitry E, Hammel P, Legoux JL, Rougier P, Bedenne L, Seitz JF (2010) Time until definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. *Eur J Cancer* **46**(15)**:** 2753-62

Boucekine M, Loundou A, Baumstarck K, Minaya-Flores P, Pelletier J, Ghattas B, Auquier P (2013) Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: a cohort study. *BMC Med Res Methodol* **13:** 20

Brady MJ, Cella DF, Mo F, Bonomi AE, Tulsky DS, Lloyd SR, Deasy S, Cobleigh M, Shiomoto G (1997) Reliability and validity of the Functional Assessment of Cancer Therapy-Breast quality-of-life instrument. *J Clin Oncol* **15**(3)**:** 974-86

Brooks MM, Jenkins LS, Schron EB, Steinberg JS, Cross JA, Paeth DS (1998) Quality of life at baseline: is assessment after randomization valid? The AVID Investigators. The Antiarrhythmics Versus Implantable Defibrillators. *Med Care* **36**(10)**:** 1515-9

Bruner DW, Hanisch LJ, Reeve BB, Trotti AM, Schrag D, Sit L, Mendoza TR, Minasian L, O'Mara A, Denicoff AM, Rowland JH, Montello M, Geoghegan C, Abernethy AP, Clauser SB, Castro K, Mitchell SA, Burke L, Trentacosti AM, Basch EM (2011) Stakeholder perspectives on implementing the National Cancer Institute's patient-reported outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). *Transl Behav Med* **1**(1)**:** 110-22

Bullinger M (2002) Assessing health related quality of life in medicine. An overview over concepts, methods and applications in international research. *Restor Neurol Neurosci* **20**(3-4)**:** 93-101

Burris HA, 3rd, Lebrun F, Rugo HS, Beck JT, Piccart M, Neven P, Baselga J, Petrakova K, Hortobagyi GN, Komorowski A, Chouinard E, Young R, Gnant M, Pritchard KI, Bennett L, Ricci JF, Bauly H, Taran T, Sahmoud T, Noguchi S (2013) Health-related quality of life of patients with advanced breast cancer treated with everolimus plus exemestane versus placebo plus exemestane in the phase 3, randomized, controlled, BOLERO-2 trial. *Cancer* **119**(10)**:** 1908-15

Buskirk TD, Stein KD (2008) Telephone vs. mail survey gives different SF-36 quality-of-life scores among cancer survivors. *Journal of clinical epidemiology* **61**(10)**:** 1049-55

Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, Group CP (2013) Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA* **309**(8)**:** 814-22

Calvert M, Blazeby J, Revicki D, Moher D, Brundage M (2011) Reporting quality of life in clinical trials: a CONSORT extension. *Lancet* **378**(9804)**:** 1684-5

Cappelleri JC, Jason Lundy J, Hays RD (2014) Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clin Ther* **36**(5)**:** 648-62

Carr AJ, Gibson B, Robinson PG (2001) Measuring quality of life: Is quality of life determined by expectations or experience? *BMJ* **322**(7296)**:** 1240-3

Cella D, Hahn EA, Dineen K (2002) Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. *Qual Life Res* **11**(3)**:** 207-21

Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, Silberman M, Yellen SB, Winicour P, Brannon J, et al. (1993) The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. *J Clin Oncol* **11**(3)**:** 570-9

Chavance M (2004) Handling missing items in quality of life studies. *Communications in Statistics-Theory and Methods* **33**(6)**:** 1371-1383

Chen LM, Ibrahim JG, Chu H (2011) Sample size and power determination in joint modeling of longitudinal and survival data. *Stat Med* **30**(18)**:** 2295-309

Cheung K, Oemar M, Oppe M, Rabin R (2009) EQ-5D user guide: basic information on how to use EQ-5D. *Rotterdam: EuroQol Group*

Cheung YB, Goh C, Thumboo J, Khoo KS, Wee J (2006) Quality of life scores differed according to mode of administration in a review of three major oncology questionnaires. *Journal of clinical epidemiology* **59**(2)**:** 185-91

Chinot OL, Wick W, Mason W, Henriksson R, Saran F, Nishikawa R, Carpentier AF, Hoang-Xuan K, Kavan P, Cernea D, Brandes AA, Hilton M, Abrey L, Cloughesy T (2014) Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. *N Engl J Med* **370**(8)**:** 709-22

Cnaan A, Laird NM, Slasor P (1997) Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat Med* **16**(20)**:** 2349-80

Cocks K, King MT, Velikova G, Martyn St-James M, Fayers PM, Brown JM (2011) Evidence-based guidelines for determination of sample size and interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. *J Clin Oncol* **29**(1)**:** 89-96

Coen JJ, Paly JJ, Niemierko A, Weyman E, Rodrigues A, Shipley WU, Zietman AL, Talcott JA (2012) Long-term quality of life outcome after proton beam monotherapy for localized prostate cancer. *Int J Radiat Oncol Biol Phys* **82**(2)**:** e201-9

Cohen J (2013) *Statistical power analysis for the behavioral sciences*: Routledge Academic

Cole BF, Bonetti M, Zaslavsky AM, Gelber RD (2005) A multistate Markov chain model for longitudinal, categorical quality-of-life data subject to non-ignorable missingness. *Stat Med* **24**(15)**:** 2317-34

Cottone F, Efficace F, Apolone G, Collins GS (2013) The added value of propensity score matching when using health-related quality of life reference data. *Stat Med* **32**(29)**:** 5119-32

Cox DR (1972) Regression models and life tables. *Journal of the Royal Statistical Society* **34:** 187-220

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* **16**(3)**:** 297-334

Crosby RD, Kolotkin RL, Williams GR (2003) Defining clinically meaningful change in health-related quality of life. *Journal of clinical epidemiology* **56**(5)**:** 395-407

Curran D, Bacchi M, Schmitz SF, Molenberghs G, Sylvester RJ (1998a) Identifying the types of missingness in quality of life data from clinical trials. *Stat Med* **17**(5-7)**:** 739-56

Curran D, Molenberghs G, Fayers PM, Machin D (1998b) Incomplete quality of life data in randomized trials: missing forms. *Stat Med* **17**(5-7)**:** 697-709

Dabakuyo TS, Guillemin F, Conroy T, Velten M, Jolly D, Mercier M, Causeret S, Cuisenier J, Graesslin O, Gauthier M, Bonnetain F (2013) Response shift effects on measuring post-operative quality of life among breast cancer patients: a multicenter cohort study. *Qual Life Res* **22**(1)**:** 1-11

De Ayala RJ (2009) The theory and practice of item response theory. New York : Guilford Press,.

de Bock E, Hardouin JB, Blanchin M, Le Neel T, Kubis G, Bonnaud-Antignac A, Dantan E, Sebille V (2013) Rasch-family models are more valuable than score-based approaches for analysing longitudinal patient-reported outcomes with missing data. *Stat Methods Med Res*

de Bock E, Hardouin JB, Blanchin M, Le Neel T, Kubis G, Sebille V (2014) Assessment of score- and Rasch-based methods for group comparison of longitudinal patient-reported outcomes with intermittent missing data (informative and non-informative). *Qual Life Res*

de Haes J, Curran D, Young T, Bottomley A, Flechtner H, Aaronson N, Blazeby J, Bjordal K, Brandberg Y, Greimel E, Maher J, Sprangers M, Cull A (2000) Quality of life evaluation in oncological clinical trials - the EORTC model. The EORTC Quality of Life Study Group. *Eur J Cancer* **36**(7)**:** 821-5

de Haes JC, van Knippenberg FC, Neijt JP (1990) Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam Symptom Checklist. *Br J Cancer* **62**(6)**:** 1034-8

Diggle P, Heagerty P, Liang K-Y, Zeger S (2002) *Analysis of longitudinal data*: Oxford University Press

Diggle P, Kenward MG (1994) Informative drop-out in longitudinal data analysis. *Applied statistics***:** 49-93

Douglas JA (1999) Item response models for longitudinal quality of life data in clinical trials. *Stat Med* **18**(21)**:** 2917-31

Doward LC, McKenna SP (2004) Defining patient-reported outcomes. *Value Health* **7 Suppl 1:** S4-8

Dunkel-Schetter C, Feinstein LG, Taylor SE, Falke RL (1992) Patterns of coping with cancer. *Health Psychology* **11**(2)**:** 79

Edelen MO, Reeve BB (2007) Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res* **16 Suppl 1:** 5-18

EuroQol G (1990) EuroQol--a new facility for the measurement of health-related quality of life. *Health policy (Amsterdam, Netherlands)* **16**(3)**:** 199

Everitt BS (2001) *Statistics for psychologists: An intermediate course*: Psychology Press

Fairclough DL (2010) *Design and analysis of quality of life studies in clinical trials*: CRC press

Fairclough DL, Fetting JH, Cella D, Wonson W, Moinpour CM (1999) Quality of life and quality adjusted survival for breast cancer patients receiving adjuvant therapy. Eastern Cooperative Oncology Group (ECOG). *Qual Life Res* **8**(8)**:** 723-31

Fairclough DL, Peterson HF, Chang V (1998) Why are missing quality of life data a problem in clinical trials of cancer therapy? *Stat Med* **17**(5-7)**:** 667-77

Fayers P, Machin D (2007) *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes*: John Wiley & Sons

Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, AobotEQoLG. B (2001) EORTC QLQ-C30 Scoring Manual (3rd edition). Brussels: EORTC 2001 ed2001.

Fayers PM, Curran D, Machin D (1998) Incomplete quality of life data in randomized trials: missing items. *Stat Med* **17**(5-7)**:** 679-96

Fayers PM, Machin D (2000) *Quality of life : Assessment, Analysis and Interpretation*: Wiley

Ferrandina G, Petrillo M, Mantegna G, Fuoco G, Terzano S, Venditti L, Marcellusi A, De Vincenzo R, Scambia G (2014) Evaluation of quality of life and emotional distress in endometrial cancer patients: A 2-year prospective, longitudinal study. *Gynecol Oncol*

Fielding S, Fayers PM, McDonald A, McPherson G, Campbell MK, Group RS (2008) Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health and quality of life outcomes* **6:** 57

Fine JP, Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**(446)**:** 496-509

Fisher GH (1995) Linear Logistic Models for Change. In *Rasch Models: Foundations, Recent Developments, and Applications*, Fisher GH, Molenaar I (eds). Springer

Fiteni F, Westeel V, Pivot X, Borg C, Vernerey D, Bonnetain F (2014) Endpoints in cancer clinical trials. *J Visc Surg* **151**(1)**:** 17-22

Floyd FJ, Widaman KF (1995) Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment* **7**(3)**:** 286

Food U, Administration D (2007) Guidance for industry: Clinical trial endpoints for the approval of cancer drugs and biologics. *Washington, DC, US Food and Drug Administration***:** 1-19

Garsa AA, Ferraro DJ, DeWees TA, Deshields TL, Margenthaler JA, Cyr AE, Naughton M, Aft R, Gillanders WE, Eberlein T, Matesa MA, Ochoa LL, Zoberi I (2013) A prospective longitudinal clinical trial evaluating quality of life after breast-conserving surgery and high-dose-rate interstitial brachytherapy for early-stage breast cancer. *Int J Radiat Oncol Biol Phys* **87**(5)**:** 1043-50

Gibbons FX (1999) Social comparison as a mediator of response shift. *Soc Sci Med* **48**(11)**:** 1517-30

Glas CA, Geerlings H, van de Laar MA, Taal E (2009) Analysis of longitudinal randomized clinical trials using item response models. *Contemp Clin Trials* **30**(2)**:** 158-70

Goel MK, Khanna P, Kishore J (2010) Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res* **1**(4)**:** 274-8

Goldhirsch A, Gelber RD, Simes RJ, Glasziou P, Coates AS (1989) Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. *J Clin Oncol* **7**(1)**:** 36-44

Gong Q, Fang L (2013) Comparison of different parametric proportional hazards models for interval-censored data: a simulation study. *Contemp Clin Trials* **36**(1)**:** 276-83

Gontero P, Oderda M, Mehnert A, Gurioli A, Marson F, Lucca I, Rink M, Schmid M, Kluth LA, Pappagallo G, Sogni F, Sanguedolce F, Schiavina R, Martorana G, Shariat SF, Chun F (2013) The impact of intravesical gemcitabine and 1/3 dose Bacillus Calmette-Guerin instillation therapy on the quality of life in patients with nonmuscle invasive bladder cancer: results of a prospective, randomized, phase II trial. *J Urol* **190**(3)**:** 857-62

Gotay CC, Korn EL, McCabe MS, Moore TD, Cheson BD (1992) Building quality of life assessment into cancer treatment studies. *Oncology (Williston Park)* **6**(6)**:** 25-8; discussion 30-2, 37

Gourgou-Bourgade S, Bascoul-Mollevi C, Desseigne F, Ychou M, Bouche O, Guimbaud R, Becouarn Y, Adenis A, Raoul JL, Boige V, Berille J, Conroy T (2013) Impact of FOLFIRINOX compared with gemcitabine on quality of life in patients with metastatic pancreatic cancer: results from the PRODIGE 4/ACCORD 11 randomized trial. *J Clin Oncol* **31**(1)**:** 23-9

Gray RJ (1988) A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics***:** 1141-1154

Groenvold M, Petersen MA, Aaronson NK, Arraras JI, Blazeby JM, Bottomley A, Fayers PM, de Graeff A, Hammerlid E, Kaasa S, Sprangers MA, Bjorner JB, Group EQoL (2006) The development of the EORTC QLQ-C15-PAL: a shortened questionnaire for cancer patients in palliative care. *Eur J Cancer* **42**(1)**:** 55-64

Guillemin F, Bombardier C, Beaton D (1993) Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *Journal of clinical epidemiology* **46**(12)**:** 1417-32

Guo Y, Logan HL, Glueck DH, Muller KE (2013) Selecting a sample size for studies with repeated measures. *BMC Med Res Methodol* **13:** 100

Hamidou Z, Dabakuyo TS, Mercier M, Fraisse J, Causeret S, Tixier H, Padeano MM, Loustalot C, Cuisenier J, Sauzedde JM, Smail M, Combier JP, Chevillote P, Rosburger C,

Arveux P, Bonnetain F (2011) Time to deterioration in quality of life score as a modality of longitudinal analysis in patients with breast cancer. *Oncologist* **16**(10)**:** 1458-68

Hardouin JB, Audureau E, Leplege A, Coste J (2012) Spatio-temporal Rasch analysis of quality of life outcomes in the French general population: measurement invariance and group comparisons. *BMC Med Res Methodol* **12:** 182

Hardouin JB, Conroy R, Sebille V (2011) Imputation by the mean score should be avoided when validating a Patient Reported Outcomes questionnaire by a Rasch model in presence of informative missing data. *BMC Med Res Methodol* **11:** 105

Hartig J, Höhler J (2009) Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation* **35**(2)**:** 57-63

Hay JL, Atkinson TM, Reeve BB, Mitchell SA, Mendoza TR, Willis G, Minasian LM, Clauser SB, Denicoff A, O'Mara A, Chen A, Bennett AV, Paul DB, Gagne J, Rogak L, Sit L, Viswanath V, Schrag D, Basch E, Group NP-CS (2014) Cognitive interviewing of the US National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). *Qual Life Res* **23**(1)**:** 257-69

He P, Kong G, Su Z (2013) Estimating the survival functions for right-censored and interval-censored data with piecewise constant hazard functions. *Contemp Clin Trials* **35**(2)**:** 122-7

Henderson R, Diggle P, Dobson A (2000) Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**(4)**:** 465-80

Heyting A, Tolboom JT, Essers JG (1992) Statistical handling of drop-outs in longitudinal clinical trials. *Stat Med* **11**(16)**:** 2043-61

Hickey AM, Bury G, O'Boyle CA, Bradley F, O'Kelly FD, Shannon W (1996) A new short form individual quality of life measure (SEIQoL-DW): application in a cohort of individuals with HIV/AIDS. *BMJ* **313**(7048)**:** 29-33

Holzner B, Giesinger JM, Pinggera J, Zugal S, Schopf F, Oberguggenberger AS, Gamper EM, Zabernigg A, Weber B, Rumpold G (2012) The Computer-based Health Evaluation

Software (CHES): a software for electronic patient-reported outcome monitoring. *BMC Med Inform Decis Mak* **12:** 126

Hong F, Bosco JL, Bush N, Berry DL (2013) Patient self-appraisal of change and minimal clinically important difference on the European organization for the research and treatment of cancer quality of life questionnaire core 30 before and during cancer therapy. *BMC Cancer* **13:** 165

Howard GS, Dailey PR, Gulianick NA (1979a) The feasibility of informed pretests in attenuating response shift bias. *Applied Psychological Measurement* **3**(3)**:** 481-494

Howard GS, Ralph KM, Gulianick NA, Maxwell SE, Nance SW, Gerber SK (1979b) Internal invalidity in pretest-posttest self-report evaluations and a reevaluation of retrospective pretests. *Applied Psychological Measurement*(3)**:** 1-23

Huisman M (2000) Imputation of missing item responses: Some simple techniques. *Quality and Quantity* **34**(4)**:** 331-351

Ibrahim JG, Chu H, Chen LM (2010) Basic concepts and methods for joint models of longitudinal and survival data. *J Clin Oncol* **28**(16)**:** 2796-801

Jaeschke R, Singer J, Guyatt GH (1989) Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled clinical trials* **10**(4)**:** 407-15

Jafari P, Bagheri Z, Ayatollahi SM, Soltani Z (2012) Using Rasch rating scale model to reassess the psychometric properties of the Persian version of the PedsQL 4.0 Generic Core Scales in school children. *Health and quality of life outcomes* **10:** 27

Jung SH, Ahn C (2003) Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Stat Med* **22**(8)**:** 1305-15

Kabbinavar FF, Wallace JF, Holmgren E, Yi J, Cella D, Yost KJ, Hurwitz HI (2008) Health-related quality of life impact of bevacizumab when combined with irinotecan, 5-fluorouracil, and leucovorin or 5-fluorouracil and leucovorin for metastatic colorectal cancer. *Oncologist* **13**(9)**:** 1021-9

Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH (1992) Comparative measurement sensitivity of short and longer health status instruments. *Med Care* **30**(10)**:** 917-25

Kazis LE, Anderson JJ, Meenan RF (1989) Effect sizes for interpreting changes in health status. *Med Care* **27**(3 Suppl)**:** S178-89

Kemmler G, Holzner B, Kopp M, Dunser M, Margreiter R, Greil R, Sperner-Unterweger B (1999) Comparison of two quality-of-life instruments for cancer patients: the functional assessment of cancer therapy-general and the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-C30. *J Clin Oncol* **17**(9)**:** 2932-40

Kim KB (2014) PFS as a surrogate for overall survival in metastatic melanoma. *Lancet Oncol* **15**(3)**:** 246-8

King-Kallimanis BL, Oort FJ, Nolte S, Schwartz CE, Sprangers MA (2011) Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients. *Qual Life Res* **20**(10)**:** 1527-40

King-Kallimanis BL, Oort FJ, Visser MR, Sprangers MA (2009) Structural equation modeling of health-related quality-of-life data illustrates the measurement and conceptual perspectives on response shift. *Journal of clinical epidemiology* **62**(11)**:** 1157-64

Kokonendji CC, Kiessé TS (2011) Discrete associated kernels method and extensions. *Statistical Methodology* **8**(6)**:** 497-516

Langer MM, Hill CD, Thissen D, Burwinkle TM, Varni JW, DeWalt DA (2008) Item response theory detected differential item functioning between healthy and ill children in quality-of-life measures. *Journal of clinical epidemiology* **61**(3)**:** 268-76

Li Y, Rapkin B (2009) Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. *Journal of clinical epidemiology* **62**(11)**:** 1138-47

Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1)**:** 13-22

Liang MH (2000) Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Med Care* **38**(9 Suppl)**:** II84-90

Lien K, Zeng L, Nguyen J, Cramarossa G, Culleton S, Caissie A, Lutz S, Chow E (2011) Comparison of the EORTC QLQ-C15-PAL and the FACIT-Pal for assessment of quality of life in patients with advanced cancer. *Expert Rev Pharmacoecon Outcomes Res* **11**(5)**:** 541-7

Littell RC, Pendergast J, Natarajan R (2000) Modelling covariance structure in the analysis of repeated measures data. *Stat Med* **19**(13)**:** 1793-819

Little RJ (1988) A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* **83**(404)**:** 1198-1202

Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA, Neaton JD, Rotnitzky A, Scharfstein D, Shih WJ, Siegel JP, Stern H (2012) The prevention and treatment of missing data in clinical trials. *N Engl J Med* **367**(14)**:** 1355-60

Little RJ, Rubin DB (1987) Statistical analysis with missing data. *New York: John Wiley & Sons*

Little RJ, Wang Y (1996) Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* **52**(1)**:** 98-111

Lix LM, Sajobi TT, Sawatzky R, Liu J, Mayo NE, Huang Y, Graff LA, Walker JR, Ediger J, Clara I, Sexton K, Carr R, Bernstein CN (2013) Relative importance measures for reprioritization response shift. *Qual Life Res* **22**(4)**:** 695-703

Lydick E, Epstein RS (1993) Interpretation of quality of life changes. *Qual Life Res* **2**(3)**:** 221-6

Lydick E, Epstein RS, Himmelberger D, White CJ (1995) Area under the curve: a metric for patient subjective responses in episodic diseases. *Qual Life Res* **4**(1)**:** 41-5

Mair P, Hatzinger R (2007a) CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science Quarterly* **49**(1)**:** 26-43

Mair P, Hatzinger R (2007b) Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software* **20**(9)**:** 1--20

Mantegna G, Petrillo M, Fuoco G, Venditti L, Terzano S, Anchora LP, Scambia G, Ferrandina G (2013) Long-term prospective longitudinal evaluation of emotional distress and quality of life in cervical cancer patients who remained disease-free 2-years from diagnosis. *BMC Cancer* **13:** 127

Marcus R, Eric P, Gabriel KR (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**(3)**:** 655-660

Mariet AS, Anota A, Maingon P, Joly F, Bosset JF, Guizard AV, Velten M, Mercier M (2014) French adaptation and psychometric validation of the Expanded Prostate cancer Index Composite questionnaire for health-related quality of life of prostate cancer patients. *ISOQOL, 21th Annual Conference, Berlin, 15-18 octobre 2014*

Masters GN (1982) A Rasch model for partial credit scoring. *Psychometrika* **47**(2)**:** 149-174

McCormack HM, Horne DJ, Sheather S (1988) Clinical applications of visual analogue scales: a critical review. *Psychol Med* **18**(4)**:** 1007-19

McHorney CA, Haley SM, Ware JE, Jr. (1997) Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *Journal of clinical epidemiology* **50**(4)**:** 451-61

McPhail S, Haines T (2010) Response shift, recall bias and their effect on measuring change in health-related quality of life amongst older hospital patients. *Health and quality of life outcomes* **8:** 65

Meiser T (2007) Rasch models for longitudinal data. In *Multivariate and Mixture Distribution Rasch Models*, pp 191-199. Springer

Mercier M, Bonneterre J, Schraub S, Lecomte S, el Hasnaoui A (1998) The development of a French version of a questionnaire on the quality of life in cancerology (Functional Living Index-Cancer: FLIC). *Bull Cancer* **85**(2)**:** 180-6

Millsap RE (2006) Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination. *Med Care* **44**(11 Suppl 3)**:** S171-5

Moinpour CM, Feigl P, Metch B, Hayden KA, Meyskens FL, Jr., Crowley J (1989) Quality of life end points in cancer clinical trials: review and recommendations. *J Natl Cancer Inst* **81**(7)**:** 485-95

Moinpour CM, Lyons B, Schmidt SP, Chansky K, Patchell RA (2000) Substituting proxy ratings for patient ratings in cancer clinical trials: an analysis based on a Southwest Oncology Group trial in patients with brain metastases. *Qual Life Res* **9**(2)**:** 219-31

Muller KE, Stewart PW (2006) Sample size for linear mixed models. *Linear Model Theory: Univariate, Multivariate, and Mixed Models***:** 385-386

Murray S, Cole B (2000) Variance and sample size calculations in quality-of-life--adjusted survival analysis (Q-TWiST). *Biometrics* **56**(1)**:** 173-82

Mustea A, Koensgen D, Belau A, Sehouli J, Lichtenegger W, Schneidewind L, Sommer H, Markmann S, Scharf JP, Ehmke M, Ledwon P, Braicu I, Zygmunt M, Koehler G (2013) Adjuvant sequential chemoradiation therapy in high-risk endometrial cancer: results of a prospective, multicenter phase-II study of the NOGGO (North-Eastern German Society of Gynaecological Oncology). *Cancer Chemother Pharmacol* **72**(5)**:** 975-83

NCI (2006) Common Terminology Criteria for Adverse Events v3.0.

Negrier S, Bushmakin AG, Cappelleri JC, Korytowsky B, Sandin R, Charbonneau C, Michaelson MD, Figlin RA, Motzer RJ (2014) Assessment of progression-free survival as a surrogate end-point for overall survival in patients with metastatic renal cell carcinoma. *Eur J Cancer* **50**(10)**:** 1766-71

Noll RB, Fairclough D (2004) Health-related quality of life: developmental and psychometric issues. *J Pediatr* **145**(1)**:** 8-9

Nolte S, Elsworth GR, Sinclair AJ, Osborne RH (2009) Tests of measurement invariance failed to support the application of the "then-test". *Journal of clinical epidemiology* **62**(11)**:** 1173-80

Norman GR, Sloan JA, Wyrwich KW (2003) Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* **41**(5)**:** 582-92

Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C (2004) Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Med Care* **42**(1 Suppl)**:** I25-36

Nunnally J, Bernstein I (1994) Psychometric Theory (3) McGraw-Hill. *New York*

O'Boyle CA, McGee H, Hickey A, O'Malley K, Joyce CR (1992) Individual quality of life in patients undergoing hip replacement. *Lancet* **339**(8801)**:** 1088-91

Oort FJ (2005) Using structural equation modeling to detect response shifts and true change. *Qual Life Res* **14**(3)**:** 587-98

Oort FJ, Visser MR, Sprangers MA (2005) An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Qual Life Res* **14**(3)**:** 599-609

Osoba D (1992) The Quality of Life Committee of the Clinical Trials Group of the National Cancer Institute of Canada: organization and functions. *Qual Life Res* **1**(3)**:** 211-8

Osoba D (2007) Translating the science of patient-reported outcomes assessment into clinical practice. *J Natl Cancer Inst Monogr*(37)**:** 5-11

Osoba D (2011) Health-related quality of life and cancer clinical trials. *Ther Adv Med Oncol* **3**(2)**:** 57-71

Osoba D, Rodrigues G, Myles J, Zee B, Pater J (1998) Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol* **16**(1)**:** 139-44

Pagano IS, Gotay CC (2005) Ethnic differential item functioning in the assessment of quality of life in cancer patients. *Health and quality of life outcomes* **3:** 60

Pan AW, Chen YL, Chung LI, Wang JD, Chen TJ, Hsiung PC (2012) A longitudinal study of the predictors of quality of life in patients with major depressive disorder utilizing a linear mixed effect model. *Psychiatry Res* **198**(3)**:** 412-9

Panageas KS, Ben-Porat L, Dickler MN, Chapman PB, Schrag D (2007) When you look matters: the effect of assessment schedule on progression-free survival. *J Natl Cancer Inst* **99**(6)**:** 428-32

Pardo Y, Guedea F, Aguilo F, Fernandez P, Macias V, Marino A, Hervas A, Herruzo I, Ortiz MJ, Ponce de Leon J, Craven-Bratle J, Suarez JF, Boladeras A, Pont A, Ayala A, Sancho G, Martinez E, Alonso J, Ferrer M (2010) Quality-of-life impact of primary treatments for localized prostate cancer in patients without hormonal treatment. *J Clin Oncol* **28**(31)**:** 4687-96

Pauler DK, McCoy S, Moinpour C (2003) Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Stat Med* **22**(5)**:** 795-809

Penar-Zadarko B, Binkowska-Bury M, Wolan M, Gawelko J, Urbanski K (2013) Longitudinal assessment of quality of life in ovarian cancer patients. *Eur J Oncol Nurs* **17**(3)**:** 381-5

Peppercorn JM, Smith TJ, Helft PR, Debono DJ, Berry SR, Wollins DS, Hayes DM, Von Roenn JH, Schnipper LE, American Society of Clinical O (2011) American society of clinical oncology statement: toward individualized care for patients with advanced cancer. *J Clin Oncol* **29**(6)**:** 755-60

Petersen MA, Groenvold M, Aaronson N, Blazeby J, Brandberg Y, de Graeff A, Fayers P, Hammerlid E, Sprangers M, Velikova G, Bjorner JB, European Organisation for R,

Treatment of Cancer Quality of Life G (2006) Item response theory was used to shorten EORTC QLQ-C30 scales for use in palliative care. *Journal of clinical epidemiology* **59**(1)**:** 36-44

Petersen MA, Groenvold M, Aaronson N, Brenne E, Fayers P, Nielsen JD, Sprangers M, Bjorner JB, European Organisation for R, Treatment of Cancer Quality of Life G (2005) Scoring based on item response theory did not alter the measurement ability of EORTC QLQ-C30 scales. *Journal of clinical epidemiology* **58**(9)**:** 902-8

Petersen MA, Groenvold M, Bjorner JB, Aaronson N, Conroy T, Cull A, Fayers P, Hjermstad M, Sprangers M, Sullivan M, European Organisation for R, Treatment of Cancer Quality of Life G (2003) Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Qual Life Res* **12**(4)**:** 373-85

Peyre H, Leplege A, Coste J (2011) Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Qual Life Res* **20**(2)**:** 287-300

Pickard AS, Knight SJ (2005) Proxy evaluation of health-related quality of life: a conceptual framework for understanding multiple proxy perspectives. *Med Care* **43**(5)**:** 493-9

Popovic M, Nguyen J, Chen E, Di Giovanni J, Zeng L, Chow E (2012) Comparison of the EORTC QLQ-BM22 and the FACT-BP for assessment of quality of life in cancer patients with bone metastases. *Expert Rev Pharmacoecon Outcomes Res* **12**(2)**:** 213-9

Post WJ, Buijs C, Stolk RP, de Vries EG, le Cessie S (2010) The analysis of longitudinal quality of life measures with informative drop-out: a pattern mixture approach. *Qual Life Res* **19**(1)**:** 137-48

Postel-Vinay S, Arkenau HT, Olmos D, Ang J, Barriuso J, Ashley S, Banerji U, De-Bono J, Judson I, Kaye S (2009) Clinical benefit in Phase-I trials of novel molecularly targeted agents: does dose matter? *Br J Cancer* **100**(9)**:** 1373-8

Pusic AL, Klassen AF, Scott AM, Klok JA, Cordeiro PG, Cano SJ (2009) Development of a new patient-reported outcome measure for breast surgery: the BREAST-Q. *Plast Reconstr Surg* **124**(2)**:** 345-53

Raboud JM, Singer J, Thorne A, Schechter MT, Shafran SD (1998) Estimating the effect of treatment on quality of life in the presence of missing data due to drop-out and death. *Qual Life Res* **7**(6)**:** 487-94

Rasch G (1993) *Probabilistic models for some intelligence and attainment tests*: ERIC

Reckase MD (2009) *Multidimensional item response theory*: Springer

Rees J, Clarke MG, Waldron D, O'Boyle C, Ewings P, MacDonagh RP (2005) The measurement of response shift in patients with advanced prostate cancer and their partners. *Health and quality of life outcomes* **3:** 21

Revicki D, Hays RD, Cella D, Sloan J (2008) Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of clinical epidemiology* **61**(2)**:** 102-9

Ring L, Hofer S, Heuston F, Harris D, O'Boyle CA (2005) Response shift masks the treatment impact on patient reported outcomes (PROs): the example of individual quality of life in edentulous patients. *Health and quality of life outcomes* **3:** 55

RJ DA (2009) The theory and practice of item response theory. New York : Guilford Press.

Rouanne M, Massard C, Hollebecque A, Rousseau V, Varga A, Gazzah A, Neuzillet Y, Lebret T, Soria JC (2013) Evaluation of sexuality, health-related quality-of-life and depression in advanced cancer patients: a prospective study in a Phase I clinical trial unit of predominantly targeted anticancer drugs. *Eur J Cancer* **49**(2)**:** 431-8

Rouquette A, Blanchin M, Sebille V, Guillemin F, Cote SM, Falissard B, Hardouin JB (2014) The minimal clinically important difference determined using item response theory models: an attempt to solve the issue of the association with baseline score. *Journal of clinical epidemiology* **67**(4)**:** 433-40

Royston P, Parmar MK (2011) The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine* **30**(19)**:** 2409-2421

Royston P, Parmar MK (2013) Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol* **13**(1)**:** 152

Samejima F (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*

Schaake W, de Groot M, Krijnen WP, Langendijk JA, van den Bergh AC (2013) Quality of life among prostate cancer patients: a prospective longitudinal population-based study. *Radiother Oncol* **108**(2)**:** 299-305

Schipper H, Clinch J, McMurray A, Levitt M (1984) Measuring the quality of life of cancer patients: the Functional Living Index-Cancer: development and validation. *J Clin Oncol* **2**(5)**:** 472-83

Schmitt N (1982) The use of analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research* **17**(3)**:** 343-358

Schoenfeld DA (1983) Sample-size formula for the proportional-hazards regression model. *Biometrics* **39**(2)**:** 499-503

Schulz KF, Altman DG, Moher D (2010) CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *Journal of clinical epidemiology* **63**(8)**:** 834-40

Schwartz CE, Sprangers MA (1999) Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc Sci Med* **48**(11)**:** 1531-48

Schwartz CE, Sprangers MA (2010) Guidelines for improving the stringency of response shift research using the thentest. *Qual Life Res* **19**(4)**:** 455-64

Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, Gundy C, Koller M, Petersen MA, Sprangers MA, Group EQoL, the Quality of Life Cross-Cultural Meta-Analysis G (2010) Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health and quality of life outcomes* **8:** 81

Scott NW, Fayers PM, Bottomley A, Aaronson NK, de Graeff A, Groenvold M, Koller M, Petersen MA, Sprangers MA (2006) Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Qual Life Res* **15**(6)**:** 1103-15; discussion 1117-20

Sebille V, Blanchin M, Guillemin F, Falissard B, Hardouin JB (2014) A simple ratio-based approach for power and sample size determination for 2-group comparison using Rasch models. *BMC Med Res Methodol* **14**(1)**:** 87

Selby PJ, Chapman JA, Etazadi-Amoli J, Dalley D, Boyd NF (1984) The development of a method for assessing the quality of life of cancer patients. *Br J Cancer* **50**(1)**:** 13-22

Shi HY, Uen YH, Yen LC, Culbertson R, Juan CH, Hou MF (2011) Two-year quality of life after breast cancer surgery: a comparison of three surgical procedures. *Eur J Surg Oncol* **37**(8)**:** 695-702

Sloan JA, Sargent DJ, Lindman J, Allmer C, Vargas-Chanes D, Creagan ET, Bonner JA, O'Connell MJ, Dalton RJ, Rowland KM, Brooks BJ, Laurie JA (2002) A new graphic for quality adjusted life years (Q-TWiST) survival analysis: the Q-TWiST plot. *Qual Life Res* **11**(1)**:** 37-45

Spitzer WO, Dobson AJ, Hall J, Chesterman E, Levi J, Shepherd R, Battista RN, Catchlove BR (1981) Measuring the quality of life of cancer patients: a concise QL-index for use by physicians. *J Chronic Dis* **34**(12)**:** 585-97

Sprangers MA, Aaronson NK (1992) The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *Journal of clinical epidemiology* **45**(7)**:** 743-60

Sprangers MA, Cull A, Bjordal K, Groenvold M, Aaronson NK (1993) The European Organization for Research and Treatment of Cancer. Approach to quality of life

assessment: guidelines for developing questionnaire modules. EORTC Study Group on Quality of Life. *Qual Life Res* **2**(4)**:** 287-95

Sprangers MA, Groenvold M, Arraras JI, Franklin J, te Velde A, Muller M, Franzini L, Williams A, de Haes HC, Hopwood P, Cull A, Aaronson NK (1996) The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: first results from a three-country field study. *J Clin Oncol* **14**(10)**:** 2756-68

Sprangers MA, Schwartz CE (1999) Integrating response shift into health-related quality of life research: a theoretical model. *Soc Sci Med* **48**(11)**:** 1507-15

Sprangers MA, te Velde A, Aaronson NK (1999a) The construction and testing of the EORTC colorectal cancer-specific quality of life questionnaire module (QLQ-CR38). European Organization for Research and Treatment of Cancer Study Group on Quality of Life. *Eur J Cancer* **35**(2)**:** 238-47

Sprangers MA, Van Dam FS, Broersen J, Lodder L, Wever L, Visser MR, Oosterveld P, Smets EM (1999b) Revealing response shift in longitudinal research on fatigue--the use of the thentest approach. *Acta Oncol* **38**(6)**:** 709-18

Steinhauser KE, Clipp EC, Bosworth HB, McNeilly M, Christakis NA, Voils CI, Tulsky JA (2004) Measuring quality of life at the end of life: validation of the QUAL-E. *Palliat Support Care* **2**(1)**:** 3-14

Stephens RJ, Hopwood P, Girling DJ, Machin D (1997) Randomized trials with quality of life endpoints: are doctors' ratings of patients' physical symptoms interchangeable with patients' self-ratings? *Qual Life Res* **6**(3)**:** 225-36

Stephenson CM, Levin RD, Spector T, Lis CG (2013) Phase I clinical trial to evaluate the safety, tolerability, and pharmacokinetics of high-dose intravenous ascorbic acid in patients with advanced cancer. *Cancer Chemother Pharmacol* **72**(1)**:** 139-46

Stockler MR, Hilpert F, Friedlander M, King MT, Wenzel L, Lee CK, Joly F, de Gregorio N, Arranz JA, Mirza MR, Sorio R, Freudensprung U, Sneller V, Hales G, Pujade-Lauraine E (2014) Patient-reported outcome results from the open-label phase III AURELIA trial evaluating bevacizumab-containing therapy for platinum-resistant ovarian cancer. *J Clin Oncol* **32**(13)**:** 1309-16

Streiner DL, Norman GR (2008) *Health measurement scales: a practical guide to their development and use*: Oxford university press

Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH (1996) Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *Journal of clinical epidemiology* **49**(7)**:** 711-7

Team RDC (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Tennant A, Conaghan PG (2007) The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and rheumatism* **57**(8)**:** 1358-62

Teresi JA (2006) Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. *Med Care* **44**(11 Suppl 3)**:** S152-70

Teresi JA, Fleishman JA (2007) Differential item functioning and health assessment. *Qual Life Res* **16 Suppl 1:** 33-42

Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, Morales LS, Orlando-Edelen M, Cella D (2007) Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. *Qual Life Res* **16 Suppl 1:** 43-68

Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, Croft P, de Vet HC (2010) Mind the MIC: large variation among populations and methods. *Journal of clinical epidemiology* **63**(5)**:** 524-34

Thijs H, Molenberghs G, Michiels B, Verbeke G, Curran D (2002) Strategies to fit pattern-mixture models. *Biostatistics* **3**(2)**:** 245-65

Trojano M, Pellegrini F, Paolicelli D, Fuiani A, Di Renzo V (2009) observational studies: propensity score analysis of non-randomized data. *Int MS J* **16**(3)**:** 90-7

van Andel G, Bottomley A, Fossa SD, Efficace F, Coens C, Guerif S, Kynaston H, Gontero P, Thalmann G, Akdas A, D'Haese S, Aaronson NK (2008) An international field study of the EORTC QLQ-PR25: a questionnaire for assessing the health-related quality of life of patients with prostate cancer. *Eur J Cancer* **44**(16)**:** 2418-24

Van der Linden WJ, Hambleton RK (1997) *Handbook of Modern Item Response Theory*, Springer Verlag, New York edn

Verweij J, Disis ML, Cannistra SA (2010) Phase I studies of drug combinations. *J Clin Oncol* **28**(30)**:** 4545-6

Wang C, Douglas J, Anderson S (2002) Item response models for joint analysis of quality of life and survival. *Stat Med* **21**(1)**:** 129-42

Ward MM, Guthrie LC, Alba M (2014) Dependence of the minimal clinically important improvement on the baseline value is a consequence of floor and ceiling effects and not different expectations by patients. *Journal of clinical epidemiology* **67**(6)**:** 689-96

Ware JE, Jr., Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* **30**(6)**:** 473-83

Wedderburn RW (1974) Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika* **61**(3)**:** 439-447

Wei JT (2002) Scoring Instructions for the Expanded Prostate cancer Index Composite (EPIC). *Ann Arbor* **1001:** 48109-0330

Wei JT, Dunn RL, Litwin MS, Sandler HM, Sanda MG (2000) Development and validation of the expanded prostate cancer index composite (EPIC) for comprehensive assessment of health-related quality of life in men with prostate cancer. *Urology* **56**(6)**:** 899-905

Weinberger M, Oddone EZ, Samsa GP, Landsman PB (1996) Are health-related quality-of-life measures affected by the mode of administration? *Journal of clinical epidemiology* **49**(2)**:** 135-40

Weis J, Arraras JI, Conroy T, Efficace F, Fleissner C, Gorog A, Hammerlid E, Holzner B, Jones L, Lanceley A, Singer S, Wirtz M, Flechtner H, Bottomley A (2013) Development of an EORTC quality of life phase III module measuring cancer-related fatigue (EORTC QLQ-FA13). *Psychooncology* **22**(5)**:** 1002-7

Wheelwright S, Darlington AS, Fitzsimmons D, Fayers P, Arraras JI, Bonnetain F, Brain E, Bredart A, Chie WC, Giesinger J, Hammerlid E, O'Connor SJ, Oerlemans S, Pallis A, Reed M, Singhal N, Vassiliou V, Young T, Johnson C (2013) International validation of the EORTC QLQ-ELD14 questionnaire for assessment of health-related quality of life elderly patients with cancer. *Br J Cancer* **109**(4)**:** 852-8

Whistance RN, Conroy T, Chie W, Costantini A, Sezer O, Koller M, Johnson CD, Pilkington SA, Arraras J, Ben-Josef E, Pullyblank AM, Fayers P, Blazeby JM, European Organisation for the R, Treatment of Cancer Quality of Life G (2009) Clinical and psychometric validation of the EORTC QLQ-CR29 questionnaire module to assess health-related quality of life in patients with colorectal cancer. *Eur J Cancer* **45**(17)**:** 3017-26

WHO (1948) WHO constitution.Geneva: WHO, 1948.

WHOQOL G (1993) Study protocol for the World Health Organization project to develop a Quality of Life assessment instrument (WHOQOL). *Quality of life Research* **2**(2)**:** 153-159

Wiklund I (2004) Assessment of patient-reported outcomes in clinical trials: the example of health-related quality of life. *Fundam Clin Pharmacol* **18**(3)**:** 351-63

Wilson IB (1999) Clinical understanding and clinical implications of response shift. *Soc Sci Med* **48**(11)**:** 1577-88

Wilson IB, Cleary PD (1995) Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA* **273**(1)**:** 59-65

Winters ZE, Balta V, Thomson HJ, Brandberg Y, Oberguggenberger A, Sinove Y, Unukovych D, Nava M, Sandelin K, Johansson H, European Organization for R, Treatment of Cancer Quality of Life G (2014) Phase III development of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire module for women undergoing breast reconstruction. *Br J Surg* **101**(4)**:** 371-82

Wu Y, Amonkar MM, Sherrill BH, O'Shaughnessy J, Ellis C, Baselga J, Blackwell KL, Burstein HJ (2011) Impact of lapatinib plus trastuzumab versus single-agent lapatinib on quality of life of patients with trastuzumab-refractory HER2+ metastatic breast cancer. *Ann Oncol* **22**(12)**:** 2582-90

Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD (1999a) Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Med Care* **37**(5)**:** 469-78

Wyrwich KW, Tierney WM, Wolinsky FD (1999b) Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *Journal of clinical epidemiology* **52**(9)**:** 861-73

Yellen SB, Cella DF, Webster K, Blendowski C, Kaplan E (1997) Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. *J Pain Symptom Manage* **13**(2)**:** 63-74

Yost KJ, Sorensen MV, Hahn EA, Glendenning GA, Gnanasakthy A, Cella D (2005) Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the Functional Assessment of Cancer Therapy-Biologic Response Modifiers (FACT-BRM) instrument. *Value Health* **8**(2)**:** 117-27

Zee BC (1998) Growth curve model analysis for quality of life data. *Stat Med* **17**(5-7)**:** 757-66

Zeger SL, Liang KY, Albert PS (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**(4)**:** 1049-60

Zumbo BD (1999) A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*

ANNEXES

Annexe A : Questionnaire de qualité de vie EORTC QLQ-C30 spécifique du cancer

QUESTIONNAIRE SUR LA QUALITE DE VIE
EORTC QLQ-C30 version 3

Nous nous intéressons à vous et à votre santé. Répondez vous-même à toutes les questions en entourant le chiffre qui correspond le mieux à votre situation. Il n'y a pas de "bonne" ou de "mauvaise" réponse. Ces informations sont strictement confidentielles.

Vos initiales : .....

Date de naissance : .....

La date d'aujourd'hui : .....

Au cours de la semaine passée	**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
1. Avez-vous des difficultés à faire certains efforts physiques pénibles comme porter un sac à provision chargé ou une valise ?	1	2	3	4
2. Avez-vous des difficultés à faire une LONGUE promenade ?	1	2	3	4
3. Avez-vous des difficultés à faire un PETIT tour dehors ?	1	2	3	4
4. Etes-vous obligée de rester au lit ou dans un fauteuil la majeure partie de la journée ?	1	2	3	4
5. Avez-vous besoin d'aide pour manger, vous habiller, faire votre toilette ou aller aux W.C. ?	1	2	3	4
6. Etes-vous limitée d'une manière ou d'une autre pour accomplir, soit votre travail, soit vos tâches habituelles chez vous ?	1	2	3	4
7. Etes-vous totalement incapable de travailler ou d'accomplir des tâches habituelles chez vous ?	1	2	3	4

Au cours de la semaine passée	**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
8. Avez-vous eu le souffle court ?	1	2	3	4
9. Avez-vous eu mal ?	1	2	3	4
10. Avez-vous eu besoin de repos ?	1	2	3	4
11. Avez-vous eu des difficultés pour dormir ?	1	2	3	4
12. Vous êtes-vous sentie faible ?	1	2	3	4
13. Avez-vous manqué d'appétit ?	1	2	3	4

14. Avez-vous eu des nausées (mal au cœur) ?	1	2	3	4
15. Avez-vous vomi ?	1	2	3	4
16. Avez-vous été constipée ?	1	2	3	4
Au cours de la semaine passée	**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
17. Avez-vous eu de la diarrhée ?	1	2	3	4
18. Etiez-vous fatiguée ?	1	2	3	4
19. Des douleurs ont-elles perturbé vos activités quotidiennes ?	1	2	3	4
20. Avez-vous eu des difficultés à vous concentrer sur certaines choses par exemple pour lire le journal ou regarder la télévision ?	1	2	3	4
21. Vous êtes-vous sentie tendue ?	1	2	3	4
22. Vous êtes-vous fait du souci ?	1	2	3	4
23. Vous êtes vous sentie irritable ?	1	2	3	4
24. Vous êtes vous sentie déprimée ?	1	2	3	4
25. Avez-vous eu des difficultés pour vous souvenir de certaines choses ?	1	2	3	4
26. Votre état physique ou votre traitement médical vous ont-ils gênée dans votre vie FAMILIALE ?	1	2	3	4
27. Votre état physique ou votre traitement médical vous ont-ils gênée dans vos activités SOCIALES (par exemple sortir avec des amis, aller au cinéma…)	1	2	3	4
28. Votre état physique ou votre traitement médical vous ont-ils causé des problèmes financiers ?	1	2	3	4

POUR LES QUESTIONS SUIVANTES, VEUILLEZ REPONDRE EN ENTOURANT LE CHIFFRE ENTRE 1 ET 7 QUI S'APPLIQUE LE MIEUX A VOTRE SITUATION.

29. Comment évalueriez-vous l'ensemble de votre ETAT PHYSIQUE au cours de la semaine passée ?

1	2	3	4	5	6	7
Très mauvais						Excellent

30. Comment évalueriez-vous l'ensemble de votre QUALITE DE VIE au cours de la semaine passée ?

Annexe B : Questionnaire de qualité de vie EORTC QLQ-BR23 spécifique du cancer du sein

FRENCH

EORTC QLQ - BR23

Les patientes rapportent parfois les symptômes ou problèmes suivants. Pourriez-vous indiquer, s'il vous plaît, si, durant la semaine passée, vous avez été affectée par l'un de ces symptômes ou problèmes. Entourez, s'il vous plaît, le chiffre qui correspond le mieux à votre situation.

Au cours de la semaine passée:	**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
31. Avez-vous eu la bouche sèche?	1	2	3	4
32. La nourriture et la boisson avaient-elles un goût inhabituel?	1	2	3	4
33. Est-ce que vos yeux étaient irrités, larmoyants ou douloureux?	1	2	3	4
34. Avez-vous perdu des cheveux?	1	2	3	4
35. Répondez à cette question uniquement si vous avez perdu des cheveux : La perte de vos cheveux vous a-t-elle contrariée?	1	2	3	4
36. Vous êtes-vous sentie malade ou souffrante?	1	2	3	4
37. Avez-vous eu des bouffées de chaleur?	1	2	3	4
38. Avez-vous eu mal à la tête?	1	2	3	4
39. Vous êtes-vous sentie moins attirante du fait de votre maladie ou de votre traitement?	1	2	3	4
40. Vous êtes vous sentie moins féminine du fait de votre maladie ou de votre traitement?	1	2	3	4
41. Avez-vous trouvé difficile de vous regarder nue?	1	2	3	4
42. Votre corps vous a-t-il déplu?	1	2	3	4
43. Vous êtes vous inquiétée de votre santé pour l'avenir?	1	2	3	4
Au cours des quatre dernières semaines:	**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
44. Dans quelle mesure vous êtes-vous intéressée à la sexualité?	1	2	3	4
45. Avez-vous eu une activité sexuelle quelconque (avec ou sans rapport)?	1	2	3	4
46. Répondez à cette question uniquement si vous avez eu une activité sexuelle: Dans quelle mesure l'activité sexuelle vous a-t-elle procuré du plaisir?	1	2	3	4

Passez à la page suivante S.V.P.

Au cours de la semaine passée:	Pas du tout	Un peu	Assez	Beaucoup
47. Avez-vous eu mal au bras ou à l'épaule?	1	2	3	4
48. Avez-vous eu la main ou le bras enflé?	1	2	3	4
49. Avez-vous eu du mal à lever le bras devant vous ou sur le côté?	1	2	3	4
50. Avez-vous ressenti des douleurs dans la région du sein traité?	1	2	3	4
51. La région de votre sein traité était-elle enflée?	1	2	3	4
52. La région de votre sein traité était-elle particulièrement sensible?	1	2	3	4
53. Avez-vous eu des problèmes de peau dans la région de votre sein traité (démangeaisons, peau qui pèle, peau sèche)?	1	2	3	4

Annexe C : Modules de QdV et de PROs validés de l'EORTC

Nom	Localisation ou spécialité	Domaine	Nombre d'items	Référence (auteur, année)
QLQ-BM22	Métastases osseuses		22	Chow et al., 2011
QLQ-BN20	Tumeurs cérébrales		20	Taphoorn, et al, 2010
QLQ-BR23	Cancer du sein		23	Sprangers et al., 1996
QLQ-CR29	Cancer colorectal		29	Whistance et al., 2009
QLQ-CX24	Cancer des cervicales		24	Greimel et al., 2006
QLQ-ELD14		Personnes âgées	14	Wheelwright et al., 2013
QLQ-EN24	Cancer de l'endomètre		24	Greimel et al., 2011
QLQ-GINET21	Tumeur neuroendocrines gastro-intestinales		21	Yadegarfar et al., 2013
QLQ-HCC18	Carcinome hépatocellulaire		18	Chie et al., 2012
QLQ-H&N35	Cancer tête et cou		35	Bjordal et al., 1999
QLQ-INFO25		Module information	25	Arraras et al., 2010
QLQ-LC13	Cancer du poumon		13	Bergman et al., 1994
QLQ-LMC21	Cancer colorectal avec métastases au foie		21	Blazeby et al., 2009
QLQ-MY20	Mélanome multiple		20	Cocks, et al., 2007
QLQ-OES18	Cancer de l'œsophage		18	Blazeby, et al., 2003
QLQ-OG25	Cancer oesophago-gastrique		25	Lagergren et al., 2007
QLQ-OV28	Cancer de l'ovaire		28	Greimel et al., 2003
QLQ-PR25	Cancer de la prostate		25	van Andel et al., 2008
QLQ-STO22	Cancer gastrique		22	Blazeby et al., 2004
IN-PATSAT32		Satisfaction des soins	32	Brédart et al., 2005
QLQ-C5PAL	Cancer en situation palliative		15	Groenvold et al., 2006

Annexe D : Modules de QdV et de PROs en cours de développement de l'EORTC

Nom	Localisation/spécialité	Domaine	Nombre d'items	Référence
Phase IV				Fiend et al., 2011
QLQ-BIL21			21	
QLQ-BLS24	Cancer superficiel du rein		24	
QLQ-CIPN20	Neuropathie périphériques induites par chimiothérapie		20	Postma et al., 2005
QLQ-FA13		fatigue	13	Weis et al., 2013
QLQ-OH17		santé bucco-dentaire	17	Hjermstad et al., 2012
QLQ-PAN26	Cancer du pancréas		26	Fitzsimmons et al., 2005
QLQ-PRT23	rectite radique		23	Halkett et al., 2010
QLQ-SWB32		bien-être spirituel	32	Lucette et al., 2014
QLQ-TC26	cancer du testicule		26	Holzner et al., 2013
Phase III complétée				
QLQ-BLM30	cancer musculaire invasif de la vessie		30	
QLQ-BrR24	reconstruction mammaire		24	Winters et al., 2013
QLQ-CLL16	leucémie lymphocytaire chronique		16	
QLQ-CML24	leucémie myéloïde chronique		24	Efficace et al., 2014
QLQ-H&N43	révision module cancer tête et cou		43	Singer et al., 2014
QLQ-HDC29	chimiothérapie à forte dose		29	Velikova et al., 2007
QLQ-OPT30	cancer ophtalmique		30	Brandberg et al., 2004
Phase III en cours				
QLQ-COMU34		communication	34	
non définie	cachexie cancéreuse			
non définie	lymphome de Hodgkin			
non définie	lymphome non Hodgkinien			
Phase I-II complétée				
QLQ-NPC42	carcinome du nasopharynx		42	
non définie	cancer de la vulve			
non définie	santé sexuelle			
Phase I-II en cours				
non définie	compression de la moelle épinière			
non définie	questionnaire basé sur les symptômes			
QLQ-THY	cancer de la thyroïde			
non définie	cancer anal			
non définie	satisfaction des soins pour les patients en ambulatoire			

Annexe E : Aide du package QoLR pour l'analyse longitudinale de la QdV

Package 'QoLR'

July 13, 2014

Version 1.0

Date 2014-07-13

Title Analysis of Health-Related Quality of Life in oncology

Author Amelie Anota

Maintainer Amelie Anota <aanota@chu-besancon.fr>

Depends R (>= 2.10.0), survival, zoo

Description To generate the scores of the EORTC QLQ-C30 questionnaire and supplementary modules and to determine the time to quality of life score deterioration in longitudinal analysis.

License GPL (>= 2.0)

NeedsCompilation no

Repository CRAN

Date/Publication 2014-07-13 13:02:36

R topics documented:

QoLR-package *Analysis of Health-Related Quality of Life in oncology*

Description

A set of functions to generate the scores of the EORTC QLQ-C30 questionnaire and supplementary modules. Two other programs to determine the time to deterioration in a Quality of Life score in longitudinal analysis with different definitions of deterioration explored.

Details

Package:	QoLR
Type:	Package
Version:	1.0
Date:	2014-07-13
License:	GPL (>=2.0)

A set of functions to generate the scores of the EORTC QLQ-C30 questionnaire, for example function 'scoring.QLQC30', and supplementary modules. Function 'TTD' to determine the time to deterioration in a Quality of Life score in longitudinal analysis and function 'TUDD' to determine the time until definitive deterioration

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Anota A. et al. Time to Health-related Quality of Life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality

of life to achieve standardization? Qual Life Res. 2013 Nov 26.

Bonnetain F. et al. Time until definitive deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. Eur J Cancer 2010, 46(5): 2753-2762.

Fayers PM. et al. The EORTC QLQC30 scoring manual. 3rd ed. Brussels: EORTC, 2001.

Hamidou Z. et al. Time to deterioration in quality of life score as a modality of longitudinal analysis in patients with breast cancer. The Oncologist 2011, 16(10):1458-1468.

Examples

```
# To generate the scores of the EORTC QLQ-C30 for the data frame dataqol1:
data(dataqol1)
scoring.QLQC30(dataqol1, items = 2:31)


# To determine the time to deterioration of 5 points at least as compared to
# the baseline score for the score "QoL" and the score "pain".
# For score "QoL", a deterioration is observed if the score decreases,
# thus, order equals to 1 for this score.
# For score "pain", a deterioration is observed if the score increases,
# thus, order equals to 2 for this score.
data(dataqol2)
ttd=TTD(dataqol2,score=c("QoL","pain"),order=c(1,2),MCID=5)
head(ttd)

# To determine the time until definitive deterioration of 5 points at least
# of the score "QoL" as compared to the baseline score with no further
# improvement of more than 5 points:
data(dataqol2)
ttd=TUDD(dataqol2,score="QoL",order=1,MCID=5)
head(ttd)
```

`dataqol1` *QLQ-C30 dataset*

Description

A data frame with the responses to the 30 items of the EORTC QLQ-C30 questionnaire for 20 patients

Usage

```
data(dataqol1)
```

Format

`id` subject identification number

`q1` item 1

`q2` item 2

`q3` item 3

`q4` item 4

`q5` item 5

`q6` item 6

`q7` item 7

`q8` item 8

`q9` item 9

`q10` item 10

`q11` item 11

`q12` item 12

`q13` item 13

`q14` item 14

`q15` item 15

`q16` item 16

`q17` item 17

`q18` item 18

`q19` item 19

`q20` item 20

`q21` item 21

`q22` item 22

`q23` item 23

`q24` item 24

`q25` item 25

`q26` item 26

`q27` item 27

`q28` item 28

`q29` item 29

`q30` item 30

See Also

`scoring.QLQC30`

dataqol2 *Longitudinal quality of life data*

Description

A data frame with 6 quality of life measures for 60 patients. The dataset is in long format.

Usage

```
data(dataqol2)
```

Format

`id` subject identification number

`time` visit number for quality of life assessment

`date` date of quality of life measurement

`QoL` score of global quality of life

`pain` score of pain

`arm` treatment arm

`death` date of death

first_pos *First positive element of a vector*

Description

A function to obtain the first positive element of a vector

Usage

```
first_pos(X)
```

Arguments

`X`	a vector

`maxi.false` *Last element of a boolean vector equals to FALSE*

Description

Report the position of the last element of a boolean vector equals to FALSE

Usage

```
maxi.false(vector)
```

Arguments

`vector`	a boolean vector

`maxi.time` *Report the highest score at each measurement time point*

Description

A function to report the highest score at each measurement time point taking into account all previous scores

Usage

```
maxi.time(vector)
```

Arguments

`vector`	A vector with quality of life scores

Value

a vector which the i-th value is equals to the maximum of the first values of the given vector until to the i-th position

Examples

```
vect=c(10,20,30,10,2,0,4,50,20)
maxi.time(vect)
```

`mini.time` *Report the lowest score at each measurement time point*

Description

A function to report the lowest score at each measurement time point taking into account all previous scores

Usage

```
mini.time(vector)
```

Arguments

`vector`	A vector with quality of life scores

Value

a vector which the i-th value is equals to the minimum of the first values of the given vector until to the i-th position

Examples

```
vect=c(10,20,30,10,2,0,4,50,20)
mini.time(vect)
```

`plotTTD` *Plot the Kaplan-Meier curve of the TTD or TUDD*

Description

A program that plot the time to deterioration curves according to the Kaplan-Meier estimation method for all patients or according to treatment arm. Additional information can be added such as the number of patients at risk and the number of the cumulative events

Usage

```
plotTTD(time, event, group = NULL, nrisk = FALSE, nevent = FALSE, group.names = NULL,
 t = NULL, info = FALSE, pos.info = NULL, xlab, ylab)
```

Arguments

time	vector equals to the time to deterioration or the time to censure
event	a dummy vector equals to 1 if the patient is deteriorated and 0 if not
group	the name of the variable corresponding to the treatment arm, only if you want survival curves according to treatment arm. Only two groups are allowed
nrisk	Boolean equals to FALSE by default. If nrisk is TRUE, then the number of patients at risk is printed under the curve at each t time point.
nevent	Boolean equals to FALSE by default. If event is TRUE, then the number of cumulative events is printed under the curve at each t time point. In that case, you must also fix nrisk to TRUE
group.names	if you want survival curves according to treatment arm, you must give the name of the treatment arms in the group.names vector
t	if nrisk is TRUE, you must give the time points to print the number of patients at risk in vector t
info	Boolean equals to FALSE by default. If two groups are given in the group vector, then the result of the Log-rank test and the Hazard ratio are added to the graph if info is TRUE
pos.info	the position of the Log-rank test and the Hazard ratio on the graph
xlab	a title for x axis
ylab	a title for y axis

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

Examples

```
# Kaplan-Meier survival curve of the Time until definitive deterioration
# of the score "QoL" with a minimal clinically important difference of 5 points
# as compared to the baseline score
#tudd1=TUDD(dataqol2, score="QoL", MCID=5,ref.init="baseline",ref.def="def1")
#ttd_1=merge(tudd1,unique(dataqol2[,c("id","arm")]))
## In the next graph, we added the number of patients at risk at time t
## and the result of the Log Rank Test and the Univariate Hazard Ratio
## of arm 2 vs. arm 1
#plotTTD(ttd_1$time.5.QoL,ttd_1$event.5.QoL,ttd_1$arm,nrisk=T,nevent=F,
#group.names=c("arm 1","arm 2"),t=seq(0,8,2),info=T,pos.info=c(6,0.8),
#xlab="time (months)",ylab="probability (%)")
```

`scoring.QLQBN20` *Scoring of the module EORTC QLQ-BN20 for brain cancer*

Description

A program that computes the scores of the module QLQ-BN20 specific to brain cancer according to the EORTC scoring manual.

Usage

```
scoring.QLQBN20(X, id="", items = 1:20)
```

Arguments

`X`	input data matrix or data frame with items of the EORTC QLQ-BN20 in columns. Missing values are inserted as `NA`.
`id`	name of the variable in the dataframe `X` corresponding to the patient identification number
`items`	a vector which indicates the positions of the 20 items, in the correct order. By default items are column 1 to 20 of `X`

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

`Y`	a data frame with the `id` variable and the score obtained for each dimension. Each score is represented by one column of `Y`. The names of the scores are those proposed in the scoring manual. If there is no `id` variable in the dataframe `X`, then `Y` only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Taphoorn, M.J., et al. (2010). An international validation study of the EORTC brain cancer module (EORTC QLQ-BN20) for assessing health-related quality of life and symptoms in brain cancer patients. European Journal of Cancer, 46(6), 1033-1040.

scoring.QLQBR23 *Scoring of the module EORTC QLQ-BR23*

Description

A program that computes the scores of the module QLQ-BR23 specific to breast cancer according to the EORTC scoring manual.

Usage

```
scoring.QLQBR23(X, id="", items = 1:23)
```

Arguments

X	input data matrix or data frame with items of the EORTC QLQ-BR23 in columns. Missing values are inserted as NA.
id	name of the variable in the dataframe X corresponding to the patient identification number
items	a vector which indicates the positions of the 23 items, in the correct order. By default items columns 1 to 23 of X

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

Y	a data frame with the id variable and the score obtained for each dimension. Each score is represented by one column of Y. The names of the scores are those proposed in the scoring manual. If there is no id variable in the dataframe X, then Y only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Sprangers MA, et al. (1996). The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: first results from a three-country field study. J Clin Oncol 14:2756-68.

`scoring.QLQC30` *Scoring of the health-related quality of life questionnaire EORTC QLQ-C30 for cancer*

Description

A program that computes the scores of the core questionnaire QLQ-C30 according to the EORTC scoring manual.

Usage

```
scoring.QLQC30(X, id="", items = 1:30)
```

Arguments

`X`	input data matrix or data frame with items of the EORTC QLQ-C30 in columns. Missing values are inserted as NA.
`id`	name of the variable in the dataframe `X` corresponding to the patient identification number
`items`	a vector which indicates the positions of the 30 items, in the correct order. By default items are columns 1 to 30 of `X`

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

`Y`	a data frame with the `id` variable and the score obtained for each dimension. Each score is represented by one column of `Y`. The names of the scores are those proposed in the scoring manual. If there is no `id` variable in the dataframe `X`, then `Y` only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Aaronson N.K., et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute, 85(5), 365-376.

Fayers PM. et al. The EORTC QLQC30 scoring manual. 3rd ed. Brussels: EORTC, 2001.

Examples

```
# scoring of the data frame dataqol1:
data(dataqol1)
scoring.QLQC30(dataqol1, id="id", items = 2:31)
```

`scoring.QLQCR29` *Scoring of the module EORTC QLQ-CR29 for colorectal cancer*

Description

A program that computes the scores of the module QLQ-CR29 according to the EORTC scoring manual.

Usage

```
scoring.QLQCR29(X, id="", items = 1:29)
```

Arguments

`X`	input data matrix or data frame with items of the EORTC QLQ-CR29 in columns. Missing values are inserted as NA.
`id`	name of the variable in the dataframe `X` corresponding to the patient identification number
`items`	a vector which indicates the positions of the 29 items, in the correct order. By default items are columns 1 to 29 of `X`

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

`Y`	a data frame with the `id` variable and the score obtained for each dimension. Each score is represented by one column of `Y`. The names of the scores are those proposed in the scoring manual. If there is no `id` variable in the dataframe `X`, then `Y` only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Whistance RN, et al. Clinical and psychometric validation of the EORTC QLQ-CR29 questionnaire module to assess health-related quality of life in patients with colorectal cancer. European journal of cancer. 2009 Nov;45(17):3017-26.

`scoring.QLQCX24` *Scoring of the module EORTC QLQ-CX24 for cervical cancer*

Description

A program that computes the scores of the module QLQ-CX24 according to the EORTC scoring manual.

Usage

```
scoring.QLQCX24(X, id="", items = 1:24)
```

Arguments

`X`	input data matrix or data frame with items of the EORTC QLQ-CX24 in columns. Missing values are inserted as NA.
`id`	name of the variable in the dataframe `X` corresponding to the patient identification number
`items`	a vector which indicates the positions of the 24 items, in the correct order. By default items are columns 1 to 24 of `X`

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

`Y`	a data frame with the `id` variable and the score obtained for each dimension. Each score is represented by one column of `Y`. The names of the scores are those proposed in the scoring manual. If there is no `id` variable in the dataframe `X`, then `Y` only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Greimel ER, et al. (2006). The European Organization for Research and Treatment of Cancer (EORTC) Quality-of-Life questionnaire cervical cancer module. Cancer, 107(8), 1812-1822.

scoring.QLQEN24 *Scoring of the module EORTC QLQ-EN24 for endometrial cancer*

Description

A program that computes the scores of the module QLQ-EN24 according to the EORTC scoring manual.

Usage

```
scoring.QLQEN24(X, id="", items = 1:24)
```

Arguments

X	input data matrix or data frame with items of the EORTC QLQ-EN24 in columns. Missing values are inserted as NA.
id	name of the variable in the dataframe X corresponding to the patient identification number
items	a vector which indicates the positions of the 24 items, in the correct order. By default items columns 1 to 24 of X

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

Y	a data frame with the id variable and the score obtained for each dimension. Each score is represented by one column of Y. The names of the scores are those proposed in the scoring manual. If there is no id variable in the dataframe X, then Y only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Greimel E., et al. (2011). Psychometric validation of the European organisation for research and treatment of cancer quality of life questionnaire-endometrial cancer module (EORTC QLQ-EN24). European Journal of Cancer, 47(2), 183-190.

scoring.QLQHN35 *Scoring of the module EORTC QLQ-H&N35*

Description

A program that computes the scores of the module QLQ-H&N35 for head and neck cancer according to the EORTC scoring manual.

Usage

```
scoring.QLQHN35(X, id="", items = 1:35)
```

Arguments

X	input data matrix or data frame with items of the EORTC QLQ-H&N35 in columns. Missing values are inserted as NA.
id	name of the variable in the dataframe X corresponding to the patient identification number
items	a vector which indicates the positions of the 35 items, in the correct order. By default items columns 1 to 35 of X

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

Y	a data frame with the id variable and the score obtained for each dimension. Each score is represented by one column of Y. The names of the scores are those proposed in the scoring manual. If there is no id variable in the dataframe X, then Y only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Bjordal, K., et al.(1999). Quality of life in head and neck cancer patients: validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-H&N35. Journal of Clinical Oncology, 17(3), 1008-1008

`scoring.QLQLC13` *Scoring of the module EORTC QLQ-LC13*

Description

A program that computes the scores of the module QLQ-LC13 for lung cancer according to the EORTC scoring manual.

Usage

```
scoring.QLQLC13(X, id="", items = 1:13)
```

Arguments

`X`	input data matrix or data frame with items of the EORTC QLQ-LC13 in columns. Missing values are inserted as NA.
`id`	name of the variable in the dataframe `X` corresponding to the patient identification number
`items`	a vector which indicates the positions of the 13 items, in the correct order. By default items columns 1 to 13 of `X`

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

`Y`	a data frame with the `id` variable and the score obtained for each dimension. Each score is represented by one column of `Y`. The names of the scores are those proposed in the scoring manual. If there is no `id` variable in the dataframe `X`, then `Y` only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Bergman B, et al. (1994). The EORTC QLQ-LC13: a modular supplement to the EORTC Core Quality of Life Questionnaire (QLQ-C30) for use in lung cancer clinical trials. EORTC Study Group on Quality of Life. Eur J Cancer 30A:635-42.

`scoring.QLQMY20` *Scoring of the module EORTC QLQ-MY20*

Description

A program that computes the scores of the module QLQ-MY20 for myeloma according to the EORTC scoring manual.

Usage

```
scoring.QLQMY20(X, id="", items = 1:20)
```

Arguments

`X`	input data matrix or data frame with items of the EORTC QLQ-MY20 in columns. Missing values are inserted as NA.
`id`	name of the variable in the dataframe `X` corresponding to the patient identification number
`items`	a vector which indicates the positions of the 20 items, in the correct order. By default items columns 1 to 20 of `X`

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

`Y`	a data frame with the `id` variable and the score obtained for each dimension. Each score is represented by one column of `Y`. The names of the scores are those proposed in the scoring manual. If there is no `id` variable in the dataframe `X`, then `Y` only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Cocks, K., et al. (2007). An international field study of the reliability and validity of a disease-specific questionnaire module (the QLQ-MY20) in assessing the quality of life of patients with multiple myeloma. European Journal of Cancer, 43(11), 1670-1678.

`scoring.QLQOES18` *Scoring of the module EORTC QLQ-OES18*

Description

A program that computes the scores of the module QLQ-OES18 for oesophageal cancer according to the EORTC scoring manual.

Usage

```
scoring.QLQOES18(X, id="", items = 1:18)
```

Arguments

`X`	input data matrix or data frame with items of the EORTC QLQ-OES18 in columns. Missing values are inserted as NA.
`id`	name of the variable in the dataframe `X` corresponding to the patient identification number
`items`	a vector which indicates the positions of the 18 items, in the correct order. By default items columns 1 to 18 of `X`

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

`Y`	a data frame with the `id` variable and the score obtained for each dimension. Each score is represented by one column of `Y`. The names of the scores are those proposed in the scoring manual. If there is no `id` variable in the dataframe `X`, then `Y` only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Blazeby, J. M., et al. (2003). Clinical and psychometric validation of an EORTC questionnaire module, the EORTC QLQ-OES18, to assess quality of life in patients with oesophageal cancer. European Journal of Cancer, 39(10), 1384-1394.

scoring.QLQOG25 *Scoring of the module EORTC QLQ-OG25*

Description

A program that computes the scores of the module QLQ-OG25 for oesophago-gastric cancer according to the EORTC scoring manual.

Usage

```
scoring.QLQOG25(X, id="", items = 1:25)
```

Arguments

X	input data matrix or data frame with items of the EORTC QLQ-OG25 in columns. Missing values are inserted as NA.
id	name of the variable in the dataframe X corresponding to the patient identification number
items	a vector which indicates the positions of the 25 items, in the correct order. By default items columns 1 to 25 of X

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

Y	a data frame with the id variable and the score obtained for each dimension. Each score is represented by one column of Y. The names of the scores are those proposed in the scoring manual. If there is no id variable in the dataframe X, then Y only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Lagergren, P., et al.(2007). Clinical and psychometric validation of a questionnaire module, the EORTC QLQ-OG25, to assess health-related quality of life in patients with cancer of the oesophagus, the oesophago-gastric junction and the stomach. European Journal of Cancer, 43(14), 2066-2073.

scoring.QLQPR25 *Scoring of the module EORTC QLQ-PR25*

Description

A program that computes the scores of the module QLQ-PR25 for prostate cancer according to the EORTC scoring manual.

Usage

```
scoring.QLQPR25(X, id="", items = 1:25)
```

Arguments

X	input data matrix or data frame with items of the EORTC QLQ-PR25 in columns. Missing values are inserted as NA.
id	name of the variable in the dataframe X corresponding to the patient identification number
items	a vector which indicates the positions of the 25 items, in the correct order. By default items columns 1 to 25 of X

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

Y	a data frame with the id variable and the score obtained for each dimension. Each score is represented by one column of Y. The names of the scores are those proposed in the scoring manual. If there is no id variable in the dataframe X, then Y only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

van Andel G, et al. An international field study of the EORTC QLQ-PR25: a questionnaire for assessing the health-related quality of life of patients with prostate cancer. European journal of cancer. 2008 Nov;44(16):2418-24.

`scoring.QLQSTO22` *Scoring of the module EORTC QLQ-STO22*

Description

A program that computes the scores of the module QLQ-STO22 for gastric cancer according to the EORTC scoring manual.

Usage

```
scoring.QLQSTO22(X, id="", items = 1:22)
```

Arguments

`X`	input data matrix or data frame with items of the EORTC QLQ-STO22 in columns. Missing values are inserted as NA.
`id`	name of the variable in the dataframe `X` corresponding to the patient identification number
`items`	a vector which indicates the positions of the 22 items, in the correct order. By default items columns 1 to 22 of `X`

Details

A score is generated if the patient answered to at least half of the corresponding items. The scores are generated according to the EORTC scoring guidelines. Simple imputation by the personal mean is retained. In this way, missing items are ignored.

Value

`Y`	a data frame with the `id` variable and the score obtained for each dimension. Each score is represented by one column of `Y`. The names of the scores are those proposed in the scoring manual. If there is no `id` variable in the dataframe `X`, then `Y` only contains the scores

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Blazeby, J.M., et al. (2004). Clinical and psychometric validation of a questionnaire module, the EORTC QLQ-STO 22, to assess quality of life in patients with gastric cancer. European Journal of Cancer, 40(15), 2260-2268.

TTD *Time to Quality of Life score deterioration*

Description

A program that computes the time to deterioration in a quality of life score.

Usage

```
TTD(X, score = "", MCID, ref.init = "baseline", order = 1,
no_baseline = "censure", no_follow = "censure", death = NA, sensitivity = FALSE)
```

Arguments

X	input data matrix or data frame with at least one quality of life score. Missing values are inserted as NA.
score	vector with the name of the quality of life scores of interest
MCID	the minimal clinically important difference
ref.init	the reference score to qualify the deterioration. By default, ref.init is "baseline", i.e. the reference score is the baseline score. If ref.init is "best", the best previous quality of life score is the reference score. If ref.init is "previous", the last previous score is the reference score.
order	a vector equals to 1 if the deterioration corresponds to a decrease of the score, 2 otherwise
no_baseline	By default, no_baseline equals to "censure" to indicate that patients with no baseline score are censured at baseline (Day 0). If no_baseline equals "event", these patients are deteriorated since baseline
no_follow	By default, no_follow equals to "censure" to indicate that patients with no follow-up score are censured just after baseline (Day 1). If no_follow equals to "event", these patients are deteriorated just after baseline
death	missing if patients who died without experienced a deterioration are censored at the time of the last quality of life assessment, equals to the name of the death date in the dataframe X otherwise
sensitivity	Boolean equals to FALSE by default. If sensitivity is TRUE, then all sensitivity analyses are performed, integrating patients with no baseline or with no follow up as event (SA1), death as event (SA2) and simultaneously no baseline, no follow and death (SA3)

Details

To apply this function, the dataset must respect a general structure. The dataset X must be in long format with the following variables in this order:

1. Patient's identification number
2. Variable identify the number of the quality of life assessment, i.e. the visit number

3. Date of quality of life measures
4. quality of life scores
5. Other variables as the date of death or the treatment arm.

The dataset must also be sorted by patient's identification number and quality of life measurement time. Dates must be in Julian format (i.e. number of days since a reference time point).

All these definitions are extensively described in the referenced papers below.

Value

The result is a dataframe with the `id` variable of the dataframe `X` and the results of the time to deterioration analyses performed.

For each score and each time to deterioration analysis, two variables are created called `event` and `time` with the name of the corresponding score as a suffix.

Moreover, if `sensitivity` is `TRUE`, a suffix is added to each result of this function reflecting the sensitivity analysis corresponding (SA1, SA2 or SA3).

The first variable `event` is a dummy vector equals to 1 if the patient is deteriorated and 0 if not. The second variable `time` equals to the time in months to deterioration since baseline if the patient is deteriorated or the time to censure.

As example, for a given score "qol" and one analyse performed (i.e. `sensitivity` is `FALSE`), then two variables are created called `event.qol` and `time.qol`.

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Anota A., et al. Time to Health-related Quality of Life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization? Qual Life Res. 2013 Nov 26.

Hamidou Z., et al. Time to deterioration in quality of life score as a modality of longitudinal analysis in patients with breast cancer. The Oncologist 2011, 16(10):1458-1468.

See Also

`TUDD`

Examples

```
data(dataqol2)
# deterioration of 5 points at least as compared to the baseline score for
# the score "QoL" and the score "pain"
TTD(dataqol2,score=c("QoL","pain"),order=1:2,MCID=5)
```

TUDD *Time until definitive deterioration in a quality of life score*

Description

A program that computes the time until definitive deterioration in quality of score.

Usage

```
TUDD(X, score = "", MCID, ref.init = "baseline", ref.def = "def1", order = 1,
 no_baseline = "censure", no_follow = "censure", death = NA, sensitivity = FALSE)
```

Arguments

X	input data matrix or data frame with a quality of life score. Missing values are inserted as NA.
score	vector with the name of the quality of life scores of interest
MCID	a vector equals to the minimal clinically important difference (MCID). Several MCID can be specified
ref.init	the reference score to qualify the deterioration. By default, ref.init is "baseline", i.e. the reference score is the baseline score. If ref.init is "best", the best previous quality of life score is the reference score. If ref.init is "previous", the last previous score is the reference score.
ref.def	the deterioration is definitive 1: if there is no clinically significant improvement as compared to the reference score ("def1"); 2: if the deterioration is also observed at all times following the deterioration ("def2"); 3: or there is no clinically significant improvement as compared to the score qualifying the deterioration ("def3")
order	a vector equals to 1 if the deterioration corresponds to a decrease of the score, 2 otherwise
no_baseline	By default, no_baseline equals to "censure" to indicate that patients with no baseline score are censured at baseline (Day 0). If no_baseline equals "event", these patients are deteriorated since baseline
no_follow	By default, no_follow equals to "censure" to indicate that patients with no follow-up score are censured just after baseline (Day 1). If no_follow equals to "event", these patients are deteriorated just after baseline
death	missing if patients who died without experienced a deterioration are censored at the time of the last quality of life assessment, equals to the name of the death date in the dataframe X otherwise
sensitivity	Boolean equals to FALSE by default. If sensitivity is TRUE, then all sensitivity analyses are performed, integrating patients with no baseline or with no follow up as event (SA1), death as event (SA2) and simultaneously no baseline, no follow and death (SA3)

Details

To apply this function, the dataset must respect a general structure. The dataset `X` must be in long format with the following variables in this order:

1. Patient's identification number
2. Variable identify the quality of life assessment, i.e. the visit number
3. Date of quality of life measure
4. quality of life scores
5. Other variables as the date of death or the treatment arm.

The dataset must also be sorted by patient's identification number and quality of life measurement time.

Dates must be in Julian format (i.e. number of days since a reference time point).

All these definitions are extensively described in the referenced papers below.

Value

The result is a dataframe with the `id` variable of the dataframe `X` and the results of the time to deterioration analyses performed.

For each score and each time to deterioration analysis, two variables are created called `event` and `time` with the value of the `MCID` and the name of the corresponding score as a suffix.

Moreover, if `sensitivity` is `TRUE`, a suffix is added to each result of this function reflecting the sensitivity analysis corresponding (SA1, SA2 or SA3).

The first variable `event` is a dummy vector equals to 1 if the patient is deteriorated and 0 if not. The second variable `time` equals to the time in months to deterioration since baseline if the patient is deteriorated or the time to censure.

As example, for a given score "qol", `MCID = 5` and one analyse performed (i.e. `sensitivity` is `FALSE`), then two variables are created called `event.5.qol` and `time.5.qol`.

Author(s)

Amelie Anota

Maintainer: Amelie Anota <aanota@chu-besancon.fr>

References

Anota A., et al. Time to Health-related Quality of Life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization? Qual Life Res. 2013 Nov 26.

Bonnetain F., et al. Time until definitive deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. Eur J Cancer 2010, 46(5): 2753-2762.

See Also

`TTD`

Examples

```
data(dataqol2)
# Time to definitive deterioration of 5 points at least of the "QoL" score
# as compared to the best previous score with no further improvement of more
# than 5 points :
ttd=TUDD(dataqol2,score=c("QoL","pain"),ref.init="best",order=1,MCID=5)
head(ttd)
```

---

whicha *To check a condition*

Description

A function to check a condition

Usage

```
whicha(condition)
```

Arguments

condition	a logical vector

Value

0 if the logical vector does not contain the value TRUE. Otherwise give the TRUE indices

---

write.TTD *Write in a csv file the results of the time to deterioration analysis*

Description

A program that computes the time to deterioration in a quality of life score and print the results in a csv file according to treatment arm

Usage

```
write.TTD(X, score = "", order = 1, ref.init = "baseline", MCID, death = NA,
group = NULL, names.group, sensitivity = TRUE, file = "")
```

Arguments

X	input data matrix or data frame with a quality of life score. Missing values are inserted as NA
score	vector with the name of the quality of life scores of interest
order	a vector equals to 1 if the deterioration corresponds to a decrease of the score, 2 otherwise
ref.init	the reference score to qualify the deterioration. By default, ref.init is "baseline", i.e. the reference score is the baseline score. If ref.init is "best", the best previous quality of life score is the reference score. If ref.init is "previous", the last previous score is the reference score.
MCID	vector equals to the minimal clinically important difference (MCID). Several MCID can be specified
death	missing if patients who died without experienced a deterioration are censored at the time of the last quality of life assessment, equals to the name of the death date in the dataframe X otherwise
group	the name of the variable in X corresponding to the treatment arm. Only two groups are allowed
names.group	the name of each treatment group to print
sensitivity	Boolean equals to TRUE by default. If sensitivity is TRUE, then all sensitivity analyses are performed, integrating patients with no baseline or with no follow up as event, death as event and simultaneously no baseline, no follow and death
file	the name of the csv file to create with the results of the time to deterioration analysis

Value

this function does not return value in R console but create a csv file with the results of the time to deterioration analysis

Author(s)

Amelie ANOTA

Maintainer: Amelie ANOTA <aanota@chu-besancon.fr>

See Also

TTD

Examples

```
### The time to deterioration of scores "QoL" and "pain" of the dataqol2 data
### set as compared to the baseline score
### with two MCID (5 points and 10 points)
### and according to the treatment arm called "arm"
### all sensitivity analyses are performed simultaneously to the main definition
### the created file is named "file_TTD_baseline.csv" and is located
```

```
### in the current directory

data(dataqol2)
write.TTD(dataqol2,score=c("QoL","pain"),order=c(1,2),MCID=c(5,10),
group="arm",names.group=c("arm 1","arm 2"),sensitivity=FALSE,
file="file_TTD_baseline")
```

`write.TUDD` *Write in a csv file the results of the time until definitive deterioration analysis*

Description

A program that computes the time until definitive deterioration in a quality of life score and print the results in a csv file according treatment arm

Usage

```
write.TUDD(X, score = "", order = 1, ref.init = "baseline", MCID, ref.def = "def1",
death = NA, group = NULL, names.group, sensitivity = TRUE, file = "")
```

Arguments

`X`	input data matrix or data frame with a quality of life score. Missing values are inserted as NA
`score`	vector with the name of the quality of life scores of interest
`order`	a vector equals to 1 if the deterioration corresponds to a decrease of the score, 2 otherwise
`ref.init`	the reference score to qualify the deterioration. By default, `ref.init` is `"baseline"`, i.e. the reference score is the baseline score. If `ref.init` is `"best"`, the best previous quality of life score is the reference score. If `ref.init` is `"previous"`, the last previous score is the reference score.
`MCID`	vector equals to the minimal clinically important difference (MCID). Several MCIDs can be specified
`ref.def`	the deterioration is definitive 1: if there is no clinically significant improvement as compared to the reference score (`"def1"`); 2: if the deterioration is also observed at all times following the deterioration (`"def1"`); 3: or there is no clinically significant improvement as compared to the score qualifying the deterioration (`"def3"`)
`death`	missing if patients who died without experienced a deterioration are censored at the time of the last QoL score, equals to the name of the death date in the dataframe `X` otherwise
`group`	the name of the variable in `X` corresponding to the treatment arm. Only two groups are allowed.
`names.group`	the name of each treatment group to print

sensitivity	Boolean equals to TRUE by default. If sensitivity is TRUE, then all sensitivity analyses are performed, integrating patients with no baseline or with no follow up as event, death as event and simultaneously no baseline, no follow and death
file	the name of the csv file to create with the results of the time to deterioration analysis

Value

this function does not return value in R console but create a csv file with the results of the time to deterioration analysis

Author(s)

Amelie ANOTA

Maintainer: Amelie ANOTA <aanota@chu-besancon.fr>

See Also

TUDD

Examples

```
### The time until definitive deterioration of scores "QoL" and "pain" of the
### dataqol2 data set as compared to the baseline score
### with two MCID (5 points and 10 points)
### and according to the treatment arm called "arm"
### all sensitivity analyses are performed simultaneously to the main definition
### the created file is named "file_TTD_baseline.csv" and is located
### in the current directory

data(dataqol2)
write.TUDD(dataqol2,score=c("QoL","pain"),order=c(1,2),MCID=c(5,10),
group="arm",names.group=c("arm 1","arm 2"),sensitivity=FALSE,
file="file_TTD_baseline")
```

Index

Résumé

La qualité de vie relative à la santé (QdV) est désormais un des objectifs majeurs des essais cliniques en cancérologie pour pouvoir s'assurer du bénéfice clinique de nouvelles stratégies thérapeutiques pour le patient. Cependant, les résultats des données de QdV restent encore peu pris en compte en pratique clinique en raison de la nature subjective et dynamique de la QdV. De plus, les méthodes statistiques pour son analyse longitudinale doivent être capables de tenir compte de l'occurrence des données manquantes et d'un potentiel effet Response Shift reflétant l'adaptation du patient vis-à-vis de la maladie et de la toxicité du traitement. Ces méthodes doivent enfin proposer des résultats facilement compréhensibles par les cliniciens.

Dans cette optique, les objectifs de ce travail ont été de faire le point sur ces facteurs limitants et de proposer des méthodes adéquates pour une interprétation robuste des données de QdV longitudinales. Ces travaux sont centrés sur la méthode du temps jusqu'à détérioration d'un score de QdV (TJD), en tant que modalité d'analyse longitudinale, ainsi que sur la caractérisation de l'occurrence de l'effet Response Shift.

Les travaux menés ont donné lieu à la création d'un package R pour l'analyse longitudinale de la QdV selon la méthode du TJD avec une interface facile d'utilisation. Certaines recommandations ont été proposées sur les définitions de TJD à appliquer selon les situations thérapeutiques et l'occurrence ou non d'un effet Response Shift. Cette méthode attractive pour les cliniciens a été appliquée dans le cadre de deux essais de phase précoces I et II. La méthode de pondération par probabilité inversée du score de propension a été investiguée conjointement avec la méthode du TJD afin de tenir compte de l'occurrence de données manquantes dépendant des caractéristiques des patients. Une comparaison de trois approches statistiques pour l'analyse longitudinale a montré la performance du modèle linéaire mixte et permet de donner quelques recommandations pour l'analyse longitudinale selon le design de l'étude. Cette étude a également montré l'impact de l'occurrence de données manquantes informatives sur les méthodes d'analyse longitudinale. Des analyses factorielles et modèles issus de la théorie de réponse à l'item ont montré leur capacité à caractériser la Response Shift conjointement avec la méthode Then-test. Enfin, bien que les modèles à équation structurelles soient régulièrement appliqués pour caractériser cet effet sur le questionnaire de QdV générique SF-36, ils semblent peu adaptés à la structure des questionnaires spécifiques du cancer du groupe « European Organization of Research and Treatment of Cancer » (EORTC).

Mots-clés : cancérologie, qualité de vie relative à la santé, analyse longitudinale, Response Shift

Abstract

Health-related quality of life (HRQoL) has become one of the major objectives of oncology clinical trials to ensure the clinical benefit of new treatment strategies for the patient. However, the results of HRQoL data remain poorly used in clinical practice due to the subjective and dynamic nature of HRQoL. Moreover, statistical methods for its longitudinal analysis have to take into account the occurrence of missing data and the potential Response Shift effect reflecting patient's adaptation of the disease and treatment toxicities. Finally, these methods should also propose some results easy understandable for clinicians.

In this context, this work aimed to review these limiting factors and to propose some suitable methods for a robust interpretation of longitudinal HRQoL data. This work is focused on both the Time to HRQoL score deterioration (TTD) as a modality of longitudinal analysis and the characterization of the occurrence of the Response Shift effect.

This work has resulted in the creation of an R package for the longitudinal HRQoL analysis according to the TTD with an easy to use interface. Some recommendations were proposed on the definitions of the TTD to apply according to the therapeutic settings and the potential occurrence of the Response Shift effect. This attractive method was applied in two early stage I and II trials. The inverse probability weighting method of the propensity score was investigated in conjunction with the TTD method to take into account the occurrence of missing data depending on patients' characteristics. A comparison between three statistical approaches for the longitudinal analysis showed the performance of the linear mixed model and allows to give some recommendations for the longitudinal analysis according to the study design. This study also highlighted the impact of the occurrence of informative missing data on the longitudinal statistical methods. Factor analyses and Item Response Theory models showed their ability to characterize the occurrence of the Response Shift in conjunction with the Then-test method. Finally, although the structural equations modeling are often used to characterize this effect on the SF-36 generic questionnaire, they seem not appropriated to the particular structure of the HRQoL cancer specific questionnaires of the European Organization of Research and Treatment of Cancer (EORTC) HRQoL group.

Keywords: oncology, health-related quality of lie, longitudinal analysis, Response Shift