UNIVERSITE PARIS SUD XI
FACULTE DE MEDECINE PARIS SUD

Année 2012 **N°**

THÈSE

en vue de l'obtention du diplôme de

Docteur de l'Université Paris Sud XI

Spécialité : Santé Publique
Option : Recherche Clinique

Ecole Doctorale de rattachement : ED 420 Santé publique

présentée et soutenue publiquement par

Caroline TOURNOUX-FACON

le 16 octobre 2012

CONTRIBUER A L'AMELIORATION DU CIBLAGE THERAPEUTIQUE EN ONCOLOGIE PAR UNE NOUVELLE METHODOLOGIE DES ESSAIS DE PHASE II

Directeur de thèse : Mme. Pascale TUBERT-BITTER
Directeur scientifique : M. Yann DE RYCKE

Composition du jury :

M. Vincent Levy	Président
M. Andrew Kramar	Rapporteur
M. Raphael Porcher	Rapporteur
M. Jean-Marc Tourani	Examinateur
Mme. Pascale Tubert-Bitter	Directeur de thèse
M. Yann De Rycke	Directeur scientifique

Remerciements :

Je tiens à remercier ma directrice de thèse Madame Pascale Tubert-Bitter de m'avoir accordé sa confiance et son soutien permanent, sa bienveillance et son aide. Je remercie également mon co-directeur de thèse Monsieur Yann De Rycke, pour sa très grande disponibilité, sa patience, ainsi que son aide précieuse. C'était un grand honneur et un vrai plaisir que d'être encadrée par eux.

Je tiens à remercier Monsieur le Professeur Vincent Levy de m'avoir fait l'honneur d'être le président de mon jury de thèse et examinateur de ce travail. Je n'oublierai pas les conseils avisés qu'il m'a donnés tout au long de ma formation.

Je remercie également mes deux rapporteurs Monsieur Raphaël Porcher, à qui je dois mes premiers programmes en statistique et Monsieur Andrew Kramar pour le temps passé, l'attention portée à ce document et leurs remarques constructives.

Je remercie aussi Monsieur le Professeur Jean-Marc Tourani d'avoir accepté de participer à ce jury de thèse et d'évaluer mon travail.

Je remercie l'équipe de biostatistiques EQ1 (ex. U780) et le service du Dr Asselain de l'Institut Curie qui m'ont accueillie chaleureusement à chaque fois que je venais de Poitiers, ainsi que Monsieur le Professeur Ingrand qui m'a permis de continuer à travailler sur cette thèse pendant mon assistanat.

Je voudrais aussi remercier Monsieur Jean Bouyer pour sa compréhension, sa confiance et son aide dans l'organisation et le bon déroulement de cette thèse.

Je souhaite remercier David et nos familles, surtout mes parents, pour leurs encouragements chaleureux, leur compréhension, l'intérêt et l'estime portés à mon travail.

Enfin, je souhaite remercier mes enfants, Elsa, qui a accepté de me voir partir à Paris toutes les semaines, Louise et Gaspard qui sont nés pendant cette thèse et qui autant que faire se peut, m'ont laissée aller jusqu'au bout.

« Ce que les parents peuvent transmettre de mieux, c'est un esprit clair, exigeant, discipliné, courageux, sans préjugé ».

Pam Brown

A mon grand-père, Louis Laurens, décédé le 05 Juillet 2012.

Résumé :

On constate que la majorité des essais de phase III, conduits après des essais de phase II pourtant prometteurs, sont “négatifs”, la nouvelle thérapeutique se révélant finalement trop toxique ou insuffisamment efficace. L’hétérogénéité de la population participant aux différentes phases de développement est une explication. Elle induirait une estimation erronée de la toxicité et, par dilution de l’effet traitement, conduirait à arrêter l’évaluation thérapeutique alors que peut être un sous-ensemble de cette population, définie à partir d’une caractéristique particulière, pourrait en bénéficier.

Dans cette thèse, nous proposons dans un premier temps une réflexion sur les aspects méthodologiques des essais de phase II qui permettraient d’améliorer l’identification précoce des thérapeutiques toxiques et des populations les plus sensibles et donc de ne planifier des essais de phase III que sur des populations encore mieux ciblées. Dans un second temps, nous présentons une nouvelle méthodologie d’essai de phase II que nous avons développée pour prendre en compte l’hétérogénéité de la population et son intérêt en pratique clinique courante. Avec cette méthode, qui est une extension du plan de Fleming à deux étapes, le développement des médicaments est moins fréquemment arrêté pour la population entière et moins de patients non sensibles à la nouvelle thérapeutique sont exposés à des molécules potentiellement toxiques, durant l’étape 2 de l’essai de phase II ou plus tard lors de l’essai de phase III.

Mots clés: Essai de phase II ; hétérogénéité de la population ; stratification ; ciblage thérapeutique

A new methodology for phase II trials to improve therapeutic targeting in oncology

Abstract :

The majority of phase III clinical trials, despite being conducted after promising phase II trials, are "negative," with the new therapy determined in the end to be too toxic or insufficiently efficacious. One explanation is the heterogeneity of the populations participating in various phases of development, which results in an erroneous estimation of the toxicity and thus a diluted therapeutic effect. This may lead to termination of evaluation of a therapy, even if a sub-population, defined by a particular characteristic, may stand to benefit from it.

In this thesis, we propose a close examination of the methodological aspects of phase II trials which would permit improved early identification of toxic therapies and of responsive populations, so that phase III trials may be designed only with the best targeted populations in mind. We present as well a new phase II clinical trial methodology which we have developed to take into account trial population heterogeneity and its importance in current clinical practice. With this method, drug development is less often stopped for the entire phase II population and less non sensitive patients are exposed to toxic drugs in the second part of phase II trials, and next in phase III trials.

Key words: phase II clinical trial; population heterogeneity; stratification; therapeutic targeting

Equipe dans laquelle cette thèse a été préparée:

Equipe Biostatistique du Centre de Recherche en Epidémiologie et Santé des Populations (CESP)

Inserm - Université Paris Sud UMRS 1018

Liste des travaux issus du travail de thèse :

Articles :

Tournoux-Facon C, De Rycke Y, Tubert-Bitter P. Targeting population entering phase III trials : a new stratified adaptive phase II design. Stat Med. 2011 Apr 15;30(8) :801-11. doi : 10.1002/sim.4148. Epub 2011 Jan 12. PubMed PMID : 21432875.

Tournoux-Facon C, De Rycke Y, Tubert-Bitter P. How a new stratified adaptive phase II design could improve targeting population. Stat Med. 2011 Jun 15;30(13) :1555-62. doi : 10.1002/sim.4211. Epub 2011 Mar 22. PubMed PMID : 21432892.

Communications affichées

C Tournoux-Facon, Y De Rycke, P Tubert-Bitter. Targeting population entering phase III trials : a new stratified adaptive phase II design. 6th International meeting on statistical methods in Biopharmacy Paris, 21-22 Sept 2009

C Tournoux-Facon, Y De Rycke, P Tubert-Bitter. A stratified adaptive phase II trial design for patient selection. 30th Annual meeting of the Society For Clinical Trials. Atlanta, 3-6 May 2009.

Table des matières

Liste des Tableaux

Liste des Figures

INTRODUCTION GENERALE

Le processus classique de développement clinique d'un nouveau traitement anti-cancéreux consiste à réaliser des essais en trois phases successives, phase I, phase II puis phase III avant toute éventuelle autorisation de mise sur le marché.

Aux études de phase I, participent des patients présentant des maladies à des stades avancés ou réfractaires pour lesquels aucun traitement standard n'est disponible. L'objectif principal de ces études est d'évaluer la sécurité et la tolérance des nouveaux traitements afin d'estimer la dose maximale recommandée pour les études de phase II. L'objectif des études de phase II est d'étudier l'activité anti-tumorale afin de déterminer si elle est cliniquement suffisante pour poursuivre l'évaluation en phase III où le traitement expérimental sera comparé au traitement de référence s'il existe ou au placebo.

Les essais de phase III incluent souvent des centaines voire des milliers de patients et peuvent se dérouler sur plusieurs années. Il a été estimé que le coût moyen d'un essai de phase III conduit par l'industrie pharmaceutique était supérieur à dix millions de dollars US. La durée moyenne des essais de phase III en oncologie a été estimée à 4.5 ans et le temps moyen dispensé par le personnel en charge de l'essai a été estimée à 200 heures par patient (Roberts, 2003 ; Emanuel, 2003). Ainsi les essais de phase III ne sont pas seulement très coûteux mais nécessitent également un travail considérable et chronophage.

Or on constate que la majorité des essais de phase III, conduits après des essais de phase II pourtant prometteurs, sont "négatifs", c'est-à-dire que le traitement expérimental se révèle finalement trop toxique ou n'est pas suffisamment efficace comparativement au traitement de référence ou au placebo (Kola, 2004 ; Roszkowski, 2000 ; Mattson, 2000 ; Fossella, 1994).

Deux questions se posent :

i) Pourquoi la toxicité qui est pourtant évaluée depuis la phase I n'a-t-elle pas été détectée avant la phase III ?
ii) N'existe-t-il pas une sous-population qui pourrait quand même bénéficier de la nouvelle thérapeutique ?

Dans les deux cas, on se retrouve face à un problème éthique :

Dans le premier, un nombre important de patients a été exposé inutilement à une thérapeutique toxique. Dans le second, cet arrêt du développement de la nouvelle thérapeutique peut constituer une réelle perte de chances pour certains patients. En effet, s'il existe en fait une sous-population aux caractéristiques particulières mal ou non identifiée au cours des essais conduits qui aurait pu bénéficier de ce traitement, alors tous les patients porteurs de cette ou de ces caractéristiques seront privés de cette nouvelle thérapeutique.

Dans ce contexte, nous avons mené une réflexion sur différents aspects méthodologiques des essais de phase II qui permettraient d'améliorer l'identification précoce des thérapeutiques toxiques et des populations les plus sensibles et donc de ne planifier des essais de phase III que sur des populations encore mieux ciblées.

Dans la première partie de cette thèse, nous détaillons le contexte méthodologique et clinique justifiant la mise en œuvre du travail. Dans les deux parties suivantes, nous présentons la nouvelle méthodologie d'essai de phase II que nous avons développée pour prendre en compte l'hétérogénéité de la population et son intérêt en pratique clinique courante. Enfin, une discussion générale portera sur la synthèse des résultats apportés et nos perspectives de travail.

CHAPITRE 1. CONTEXTES METHODOLOGIQUE ET CLINIQUE

Dans cette première partie, nous détaillons le contexte méthodologique classique des essais de phase II puis expliquons en quoi l'hétérogénéité des patients inclus non prise en compte dans les méthodes classiques des essais cliniques en oncologie peut conduire à exposer inutilement certains patients à des thérapeutiques toxiques ou à les priver d'une thérapeutique potentiellement efficace pour certains d'entre eux. Enfin nous présentons une synthèse bibliographique des schémas de phase II visant à réduire les échecs :

- en phase III pour toxicité grâce à la prise en compte conjointe de l'efficacité et de la toxicité
- en phase II pour inefficacité grâce à la prise en compte de l'hétérogénéité de la population qui améliore la sélection des patients à inclure en phase III.

METHODOLOGIE CLASSIQUE DES ESSAIS DE PHASE II EN CANCEROLOGIE

Il existe une très grande variété d'essais de phase II et plus d'une centaine de schémas publiés dans la littérature, allant du premier schéma publié en 1961 par Gehan (Gehan, 1961), à des schémas plus complexes, multi-étapes ou multi-bras proposés ces dernières années.

Le schéma le plus simple est l'essai de phase II ne comportant qu'un seul bras dans lequel tous les patients reçoivent le traitement expérimental. Cet essai peut être conduit en une ou plusieurs étapes. Ce dernier prévoit alors la possibilité d'un arrêt précoce pour inefficacité ou efficacité. Les données d'efficacité sont en général résumées par un critère, le plus classique étant le taux de réponse, mesurée à un temps donné en fonction de la réduction uni ou bi dimensionnelle de la taille de la tumeur (ou critère RECIST : Response Evaluation Criteria in Solid Tumours (Therasse, 2000)).

Les schémas multi-étapes les plus connus et utilisés sont sans doute les méthodes fréquentistes développées par Simon (Simon, 1989) ou Fleming (Fleming, 1982). Si les calculs permettant de déterminer les bornes d'arrêt à partir des niveaux de risque d'erreur consentis sont relativement complexes, les règles d'arrêt en elles-mêmes sont assez simples. Prenons l'exemple de la méthode proposée par Fleming (à deux étapes pour simplification) :
L'essai commence par étudier l'activité sur un nombre limité de patients. La détermination des effectifs à inclure à la première et éventuellement à la deuxième étapes sont calculés préalablement pour satisfaire aux risques d'erreurs consentis et donc connus par l'investigateur dès le début de l'essai. Si le nombre de succès est inférieur à la première borne d'arrêt calculée, alors l'essai est définitivement interrompu pour inefficacité. Si le nombre de succès est supérieur à la seconde borne d'arrêt calculée, alors l'essai est définitivement interrompu pour efficacité. Si le nombre de succès est

compris entre ces deux bornes d'arrêt, un groupe supplémentaire de patients est inclus lors d'une deuxième étape. A nouveau, le nombre total de réponses observées, comparé à une seule et unique borne d'arrêt également calculée initialement, conditionne la poursuite ou l'arrêt de l'évaluation du traitement : si le nombre de patients traités avec succès est suffisant, une phase III est envisagée.

Des méthodes inspirées de l'approche bayésienne ont également été proposées parmi lesquelles on peut citer celles de Thall et Simon (Thall, 1994) d'une part, et celle de Thall (Thall, 1995) d'autre part. D'autres auteurs tels que Cressie (Cressie, 1994), Heitjan (Heitjan, 1997), ou Stallard (Stallard, 1998, 1999) ont également développé des schémas d'études selon cette approche.
Un modèle bayésien comporte deux composantes principales. La première consiste comme pour les approches fréquentistes à choisir un modèle statistique pour décrire le phénomène à l'étude et donc le choix d'un paramètre d'intérêt (survie médiane, taux de réponse, …). Contrairement aux approches fréquentistes, ce paramètre est considéré comme aléatoire et présente donc une distribution *a priori* qu'il faut aussi choisir. A travers la fonction de vraisemblance, le théorème de Bayes permet alors de calculer la distribution *a posteriori* du paramètre en fonction des données observées. Dans les essais de phase II séquentiels, le théorème de Bayes est appliqué de façon répétée, en considérant chaque paramètre *a posteriori* comme le paramètre *a priori* de l'étape suivante. La méthode bayésienne est donc une approche par apprentissage, dans laquelle l'observation des données recueillies au cours d'un essai permet d'affiner la connaissance du paramètre défini *a priori* et donc l'estimation *a posteriori* du paramètre étudié. On peut reprocher à l'approche bayésienne le caractère subjectif de l'information présente dans la distribution *a priori* puisque lorsque l'on décide de faire un essai thérapeutique, c'est justement parce que l'on souhaite obtenir une réponse objective. Il est cependant des situations où ce type de méthode peut être séduisant. C'est par exemple le cas d'essais thérapeutiques portant sur des maladies très rares. Du fait de la petitesse des effectifs, un essai

habituel a peu de chance d'être concluant. Dans un tel cas, l'opinion d'experts est une source d'information essentielle, et l'approche bayésienne permet d'en tenir compte. A noter également que des analyses de sensibilité portant sur la définition du paramètre *a priori* permettent d'affiner les résultats obtenus.

Enfin, lors des essais de phase II, l'activité de la nouvelle thérapeutique n'est habituellement pas comparée aux traitements de référence s'ils existent. La supériorité, l'équivalence ou la non-infériorité n'est recherchée qu'en phase III. Cependant certains auteurs insistent sur l'intérêt de la randomisation en phase II (Estey, 2003; Sargent, 2005) qui prévoit l'allocation des patients à un ou différents bras de traitements expérimentaux ou de référence. En effet, dans le cadre d'un essai à un seul bras, l'effet observé est une combinaison de « l'effet traitement » et de « l'effet essai » (lié aux centres, aux professionnels de santé, aux soins de support…). L'effet propre du traitement est donc difficilement identifiable. Cette question de la randomisation est donc particulièrement intéressante quand l'objectif est de sélectionner en phase II parmi plusieurs traitements expérimentaux les plus prometteurs. Ces schémas comprennent plusieurs bras expérimentaux et permettent d'effectuer une sélection des traitements en éliminant les traitements trop peu efficaces ou trop toxiques au cours d'une seule étude. Ces essais sont plus complexes à mettre en œuvre. Les schémas proposés par exemple par Thall (Thal, 1989; 1993; 1998) et Schaid (Schaid, 1990) permettent une sélection des traitements à la fin de la première étape et une comparaison des traitements retenus à la fin de la seconde étape (du fait de cette comparaison on peut se demander si l'on ne sort pas du cadre stricto-sensu des essais de phase II).

C'est en phase III que la randomisation est cruciale puisque c'est à cette étape de l'évaluation d'un nouveau traitement que celui-ci va devoir démontrer son bénéfice par rapport aux traitements de référence ou à défaut au placebo. Il est donc essentiel à ce niveau de tout mettre en œuvre pour ne mesurer que l'effet traitement. C'est au terme de ces essais de phase III que

les autorisations de mise sur le marché sont susceptibles d'être données par les agences du médicament.

Il n'existe donc pas un seul schéma pour conduire les essais de phase II mais autant de méthodologies différentes que d'objectifs, qu'ils soient exploratoires permettant de générer des hypothèses sur les thérapeutiques les plus prometteuses en phase III ou confirmatoires, permettant d'affirmer l'activité d'une nouvelle thérapeutique. Or quelle que soit l'approche en phase II, fréquentiste ou bayésienne, randomisée ou non, on constate que les résultats prometteurs en phase II ne se concrétisent malheureusement pas suffisamment en phase III, les thérapeutiques étant jugées inefficaces ou finalement trop toxiques. Dans une revue des essais de phase III par exemple, conduits en Amérique du nord entre 1972 et 1990 sur des patients atteints de cancer du poumon avancé à petites cellules, seulement 5 des 21 études (24%) montraient un bénéfice statistiquement significatif du traitement expérimental (Roszkowski, 2000). La proportion d'études "positives" serait également très faible (15%) parmi les essais conduits dans le cancer non à petites cellules (Mattson, 2000 ; Fossella, 1994). Alors peut-on proposer une explication à ce fort taux d'échecs ainsi que des solutions d'amélioration ?

HETEROGENEITE DES PATIENTS INCLUS COMME SOURCE POSSIBLE D'ECHEC DES ESSAIS CLINIQUES EN CANCEROLOGIE

L'hétérogénéité clinique ou biologique des populations (liée au sexe, à l'âge, au poids, au stade de la maladie, aux antécédents, aux traitements antérieurs, à la positivité de certains biomarqueurs…) est souvent suspectée lors des échecs des essais cliniques de phase II ou III pour toxicité.

En effet, en phase I les nouveaux traitements sont généralement proposés aux patients lorsqu'ils représentent une alternative thérapeutique justifiée, ou lorsque les traitements conventionnels n'ont pas été efficaces (l'historique des traitements déjà reçus est donc assez lourd), mais aussi sous conditions d'éligibilité, notamment : un bon état général (PS 0 ou 1), des fonctions cardiaques, rénales, hépatiques et hématologiques préservées, l'absence de comorbidités majeures. Or parce que les essais de phase I sont souvent des essais de petite taille d'une part et parce que la population participant à la phase II présente des caractéristiques cliniques souvent différentes de celle de la phase I, plus proches des conditions d'utilisation envisagée par la suite d'autre part, la dose maximale tolérée en phase I peut ne pas être précisément établie pour la population des essais de phase II et III. Des thérapeutiques *a priori* bien tolérées en phase I peuvent donc se révéler toxiques lorsqu'elles sont utilisées sur de plus larges effectifs, en phase II et surtout en phase III et en particulier dans certaines situations, comme par exemple :

- Chez les sujets âgés car ils sont souvent particulièrement susceptibles de présenter des toxicités multiples du fait de leur parcours thérapeutique antérieur et de leurs comorbidités.
- Chez les enfants car la population pédiatrique des essais de phase II est souvent hétérogène du fait de l'âge.

L'hétérogénéité clinique ou biologique des populations entraînant ici l'identification éventuelle de toxicités inattendues après la phase I est également souvent suspectée lors des échecs pour inefficacité des essais de phase III. L'hypothèse d'une dilution de l'effet du traitement est régulièrement mise en avant pour justifier des analyses *post-hoc* en sous-groupes réalisées, publiées et soumises dans les dossiers de demande d'autorisation de mise sur le marché. Dans un essai qui n'a pas montré de différence statistiquement significative, le but des analyses en sous-groupes est de rechercher le ou les sous-groupes de patients dans lesquels existerait un effet du traitement statistiquement significatif. L'idée est de dire que l'effet du traitement n'existe pas chez tous les types de patients mais seulement chez certains d'entre eux. Le mélange de patients sensibles au traitement avec d'autres patients non sensibles au traitement, conduit, au niveau de l'essai, à une dilution de l'effet et à l'absence de différence significative. Pocock a revu 50 essais publiés dans des revues internationales à comité de lecture en 1997 et a noté que 70% d'entre eux ont rapporté une médiane de 4 analyses en sous-groupes (Pocock, 2002). Or, les analyses en sous-groupes se heurtent à plusieurs difficultés méthodologiques : répétitions des tests statistiques, inflation du risque alpha, perte de puissance et démarche exploratoire (Yusuf, 1991; Sleight, 2000). Elles ne sont pas pour autant totalement dénuées d'intérêt. Elles génèrent de nouvelles hypothèses, qui seront vérifiées dans de nouvelles études de confirmation. Le résultat d'une analyse en sous-groupes est donc de nature exploratoire.

Ainsi dans plusieurs études l'hétérogénéité de la population s'est traduite par une probabilité de réponse effectivement hétérogène entre des sous-groupes de patients.

Dans certains cancers, la réponse aux traitements précédents est un indicateur de réponse aux traitements ultérieurs. Dans un essai de phase III incluant 188 patients atteints de lymphomes non hodgkiniens (Guglielmi, 1998), par exemple, il a été démontré que 77 patients qui avaient rechuté dans les 12 mois suivant le diagnostic initial avaient un taux de réponse de 40 % au traitement ultérieur. Au contraire, les patients ayant rechuté après

12 mois avaient un taux de réponse de 69 % ($p < 0.001$). Deux études de patientes avec cancer des ovaires avancé (Gore, 1990 et Blackledge, 1989) ont constaté que la durée de l'intervalle libre sans maladie avant rechute était prédictive de la réponse. Dans la première étude, seules 17 % (6/35) des patientes ayant rechuté dans les 18 mois ont répondu contre 53 % (10/19) parmi les patientes ayant rechuté après 18 mois ($p < 0.01$). Les sites métastatiques de certains cancers peuvent induire également des sensibilités différentes aux traitements. Ainsi dans une étude sur le cancer de l'ovaire stade IV (Bonnefoi, 1999), 45% des patientes avec métastases pulmonaires ou hépatiques répondaient à la première ligne de traitement par chimiothérapie contre 62% des patients avec un autre site métastatique ($p<0.001$).
Au-delà du rôle prédictif de l'intervalle libre sans maladie sur la réponse tumorale en oncologie, des marqueurs biologiques de la sensibilité au traitement existent également puisqu'ils interviennent dans le mécanisme d'action du traitement. Dans ce cas, les taux de réponse vont dépendre du niveau du biomarqueur. Aschele et al. (Aschele, 1999) ont par exemple montré que les niveaux de Thymidylate synthétase étaient prédictifs de la réponse aux chimiothérapies à base de fluorouracil chez les patients atteints de cancer colorectal avancé. Les taux de réponse étaient de 67% et 24% chez les patients avec bas et haut niveaux respectivement ($p<0.01$). Dans un essai de phase III dans le cancer du poumon à petites cellules, Dingemans (Dingemans, 1999) a également retrouvé que les niveaux élevés d'expression du gène de la Topoisomérase II beta, enzyme impliquée dans la résistance aux chimiothérapies, étaient prédictifs d'un faible taux de réponse. Ces biomarqueurs identifient donc les sous-groupes de patients les mieux à même de bénéficier des nouvelles thérapeutiques.

Depuis 2000, plusieurs thérapies ont reçu une autorisation de mise sur le marché (AMM) restreinte à un sous-groupe de patients présentant des altérations moléculaires spécifiques (Tableau 1).
Ainsi, dans la prise en charge des patients atteints de cancer colorectal métastatique, plusieurs études avaient montré que seuls les patients dont la tumeur ne présentait pas de mutation du gène KRAS étaient susceptibles de

bénéficier d'un traitement par cetuximab et panitumumab (Karapetis, 2008. Van Cutsem, 2008). Dans ce contexte, l'Agence européenne du médicament a autorisé en 2007 et 2008 l'utilisation ciblée de ces traitements, en les réservant aux patients dont la tumeur porte la forme non mutée du gène KRAS.

Tableau 1. ***Liste des thérapies ciblées en cancérologie et autorisation de mise sur le marché (AMM de l'Agence européenne du médicament)***

Biomarqueur	Pathologie	Molécule	Date de l'AMM
Translocation de BCR-AB	Leucémie myéloïde chronique Leucémie aiguë lymphoblastique	Imatinib Dasatinib Nilotinib	2001 2006/2010 2007/2010
Mutations de KIT et de PDGFRA	Tumeurs stromales gastro-intestinales (GIST)	Imatinib Lapatinib	2002 2010
Amplification de HER2	Cancer du sein	Trastuzumab Lapatinib	2000 2008
Amplification de HER2	Cancer de l'estomac	Trastuzumab	2009
Mutations de KRAS	Cancer colorectal	Panitumumab Cetuximab	2007 2008
Mutations d'EGFR	Cancer du poumon	Gefitinib Erlotinib	2009 2011

Un autre exemple où l'identification de l'hétérogénéité de la population a conduit à une restriction de l'indication thérapeutique est donné par l'historique de développement des traitements des patients atteints d'une forme avancée ou métastatique de cancer du poumon. Deux inhibiteurs réversibles spécifiques de l'activité tyrosine kinase de l'EGFR (TKI-EGFR) ont été développés dans le cancer du poumon : i) le gefitinib (Iressa®), ii) l'erlotinib (Tarceva®). Ces deux inhibiteurs ont une bonne efficacité tumorale, mais celle-ci est limitée à certains patients : en population non sélectionnée, une réponse clinique est observée chez environ 10 % des patients d'origine caucasienne et chez 30 % des patients d'origine asiatique (Gazdar, 2009 ; Thatcher, 2005 ; Shepherd, 2005). Le séquençage du gène EGFR dans le tissu tumoral a montré que la majorité des tumeurs qui répondaient à ces inhibiteurs portaient des mutations dans le domaine tyrosine kinase de

l'EGFR, conduisant à une activation du récepteur et à une sensibilité accrue aux inhibiteurs de tyrosine kinase (Lynch, 2004 ; Paez, 2004 ; Pao, 2004). Ces études rétrospectives, menées sur un petit nombre de patients, tendaient à montrer que la présence d'une mutation de l'EGFR dans la tumeur des patients atteints de cancer du poumon était un facteur prédictif de la réponse aux TKI-EGFR. Suite à ces études préliminaires, plusieurs études cliniques ont analysé l'impact des mutations de l'EGFR sur la réponse, la survie sans progression et la survie des patients traités par erlotinib ou gefitinib (Tableau 2). Toutes ces études sont convergentes : les patients présentant une mutation activatrice de l'EGFR dans leur tumeur ont un bénéfice accru au traitement par TKI-EGFR en termes de taux de réponse, de durée de survie sans progression et de survie globale par rapport à ceux dont la tumeur présente la forme sauvage de l'EGFR.
Dans ce contexte, l'autorisation de mise sur le marché de ces thérapeutiques en 2009 et 2011 a été restreinte à cette sous-population des patients EGFR+.

Le ciblage thérapeutique consiste à personnaliser un traitement en fonction des caractéristiques spécifiques du malade et de sa pathologie pour une meilleure efficacité et une meilleure tolérance du traitement. Pour que cette nouvelle approche thérapeutique soit possible en pratique clinique courante, il faut pouvoir déterminer quelles sont ces caractéristiques initiales du patient ou de sa tumeur qui pour un traitement donné permettent d'envisager le meilleur résultat possible.
Les méthodologies développées pour la planification des essais cliniques doivent tenir compte de cette nouvelle approche thérapeutique et permettre dès les premières phases de l'évaluation des nouvelles thérapeutiques de sélectionner l'ensemble des traitements efficaces et tolérés, mais également d'identifier précisément l'ensemble des patients susceptibles de pouvoir en bénéficier.
L'historique du développement de certaines thérapeutiques illustre bien cette nécessité de développer de nouvelles approches méthodologiques.

Tableau 2. Réponse à l'erlotinib ou au gefitinib selon le statut mutationnel de l'EGFR

ETUDE	JACKMAN 2009	CADRANEL 2009	ROSELL 2009	MORITA 2009
Type d'étude	Méta-analyse Europe + Etats-Unis de 5 essais de phase II	Cohorte prospective	Cohorte prospective	Méta-analyse de 7 essais de phase II au Japon
Nombre de patients	317	304	2105	148
Traitement	Erlotinib Gefitinib	Erlotinib	Erlotinib	Erlotinib
Ligne de prescription	Première ligne	Première ligne et plus	Première ligne Seconde ligne	Première ligne Seconde ligne
Patients EGFR+				
Nombre de patients (%)	84 sur 2223 (38%)	43 sur 304 (14%)	350 sur 2105 (16%)	148
Taux de réponse	67%	nd	70%	76%
Durée médiane de survie sans progression	11.8 mois	8.4 mois	14 mois* 14 mois	10.7 mois* 6 mois (p<0.001)
Durée médiane de survie	23.9 mois	14.4 mois	28 mois* 27 mois	24.3 mois
Patients EGFR-				
Nombre de patients (%)	139	261	nd	nd
Taux de réponse	0-5%	nd	nd	nd
Durée médiane de survie sans progression	< 4 mois	2.3 mois	nd	nd
Durée médiane de survie	12 mois	5.5 mois	nd	nd

** Première ligne versus seconde ligne*

METHODES DE PHASE II EXISTANTES CONTRIBUANT A AMELIORER LE CIBLAGE THERAPEUTIQUE EN CANCEROLOGIE

De nombreuses molécules testées en recherche clinique ne sont jamais mises sur le marché. La molécule est jugée soit trop toxique au-delà de la phase I (en phase II voire en phase III), soit inefficace dès la phase II ou seulement en phase III.

Deux questions sont alors soulevées : la toxicité aurait-elle pu être mieux évaluée en phase II et l'arrêt complet du développement est-il vraiment justifié, n'existe-t-il pas une sous-population qui pourrait bénéficier de la nouvelle thérapeutique ?

Ces deux situations ont fait l'objet d'une recherche méthodologique et les schémas de phase II innovants proposés dans la littérature sont décrits dans ce paragraphe.

METHODOLOGIES BI-VARIEES PERMETTANT DE PRENDRE EN COMPTE CONJOINTEMENT L'EFFICACITE ET LA TOXICITE

Si le critère d'évaluation principal est l'efficacité, il est habituel, même en phase II, de surveiller la tolérance de la thérapeutique expérimentale tout au long de l'essai. Il y a classiquement deux possibilités. La première consiste simplement à fixer une règle d'arrêt indépendante des données d'efficacité, permettant de conclure que la nouvelle thérapeutique est trop toxique pour que son évaluation soit poursuivie si un niveau de toxicité jugé inacceptable est dépassé. La seconde consiste à utiliser une statistique basée sur l'évaluation d'un critère bivarié. Ainsi certains auteurs ont proposé de considérer conjointement l'activité et la toxicité d'un nouveau traitement, de telle sorte que le traitement investigué en phase II ait à démontrer à la fois son activité et son innocuité pour justifier de la poursuite de son évaluation

en phase III. Plus d'une dizaine de méthodes permettant un arrêt précoce pour toxicité ou inefficacité ont été développées depuis 1995 ; Elles sont d'inspiration fréquentiste ou bayésienne, sur critère binaire, ordinal ou de survie, à une ou plusieurs étapes. Une bibliographie exhaustive a été publiée par Sarah Brown en 2011 (Brown et al, 2011), que nous avons synthétisée en partie dans le Tableau 3 et actualisée.

Tableau 3. ***Liste des méthodes permettant la prise en compte de la toxicité dans les essais de phase II (adaptée de la revue bibliographique de S Brown)***

	Critère binaire		Critère ordinal	Critère de survie	
	Avec bras contrôle	**Sans bras contrôle**	**Avec bras contrôle**	**Avec bras contrôle**	**Sans bras contrôle**
1 étape	Thall & Cheng (1999) Boo & Zielhuis (2004)	Conaway & Petroni (1995,1996) Jin (2007)		Thall & Cheng (1999)	
2 étapes	Thall & Cheng (2001)	Bryant & Day (1995) Conaway & Petroni (1996) Thall & Cheng (2001) Jin (2007) Wu & Liu (2007) Chen & Chi (2011)	Sun (2009)	Thall & Cheng (2001)	Thall & Cheng (2001)
Multi-étapes	Thall & Cheng (2001)	Conaway & Petroni (1995) Thall (1996) Thall & Sung (1998) Thall & Cheng (2001) Thall & al (2003)		Thall & Cheng (2001)	Thall & Cheng (2001) Thall et al (2003)
Surveillance en continue		Thall (1996) Thall & Sung (1998) Thall et al (2003) Goldman (1987) Goldman & Hannan (2001) Ivanova (2005) Ray (2012)			Thall et al (2003)

Parmi les méthodes les plus innovantes, on peut citer Conaway et Petroni (Conaway & Petroni, 1995) qui proposent en 1995 un schéma à plusieurs étapes, sans bras contrôle, basé sur un critère bivarié où la réponse et la toxicité sont des variables binaires. L'investigateur doit spécifier les hypothèses nulles pour les deux paramètres, l'association entre les deux (représentée par l'odds-ratio), les erreurs de type I et II. Le nombre de sujets nécessaire est calculé par itérations pour satisfaire l'ensemble des hypothèses ci-dessus. A la fin de la première étape, l'essai peut être arrêté précocement pour manque d'efficacité ou pour toxicité inacceptable. L'extension publiée en 1996 propose des seuils permettant d'accepter une toxicité plus grande pour une efficacité maximale ou au contraire une efficacité moindre pour une toxicité nulle (Conaway & Petroni, 1996).

Une autre méthode basée sur un critère binaire a été proposée par Bryant et Day (Bryant & Day, 1995). Il s'agit d'une extension du plan de Simon à un seul bras, en deux étapes qui prend en considération la toxicité et l'efficacité. Un arrêt précoce pour inefficacité ou toxicité est possible. Le point de divergence le plus important avec la méthode proposée par Conoway et Petroni est l'hypothèse d'indépendance entre les deux critères. Dans un article publié en 2007 dans la revue Contemporary Clinical Trials (Tournoux et al, 2007) et présenté en annexe 1, nous avons lors d'un précédent travail de recherche étudié l'impact d'une erreur de cette hypothèse, sur la puissance et l'erreur de type I. Quand l'hypothèse d'indépendance est posée (comme c'est le cas avec la méthode de Bryant & Day) alors qu'on est en présence d'une association non nulle entre l'efficacité et la tolérance (efficacité et tolérance ne sont donc pas indépendantes), on observe seulement une très faible diminution de la puissance et une très faible augmentation de l'erreur de type I. A l'inverse, quand l'hypothèse d'une relation entre efficacité et tolérance est posée dès le départ (comme c'est le cas avec la méthode de Conaway & Petroni), les erreurs sur cette hypothèse ont un impact important puisque la puissance baisse et le risque d'erreur de type I augmente sensiblement. On a donc pu constater qu'il n'y avait aucun intérêt à porter une autre hypothèse que l'indépendance. Nous recommandions donc

l'utilisation de la méthode développée par Bryant et Day en 1995 qui fait l'hypothèse d'indépendance entre les deux critères.

En 2011, Chen et Chi (Chen & Chi, 2011) ont proposé une extension de la méthode de Conaway et Petroni, basée sur la procédure d'échantillonnage proposée par Phatak et Bhatt en 1967 (Phatak & Bhatt, 1967), permettant de réduire les effectifs inclus en cas d'inefficacité et/ou de toxicité et dans laquelle l'impact d'une erreur sur l'hypothèse d'indépendance est minime.

Peter Thall a également beaucoup publié sur cette question (Thall, 1996, 1998, 1999, 2001, 2003). Les schémas bayésiens développés permettent soit de choisir un traitement parmi d'autres de telle sorte que le traitement présentant la plus grande efficacité est correctement sélectionné avec une probabilité minimale préalablement déterminée, un arrêt précoce pour inefficacité ou toxicité étant possible (Un logiciel est disponible sur : http://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=3); soit d'évaluer conjointement l'efficacité et la toxicité. L'effet du traitement est basé sur la détermination de seuils acceptables de la balance bénéfice-risque et de la relation entre efficacité et toxicité décrite par l'odds-ratio (OR) où l'indépendance entre ces deux critères se traduit par OR=1 (aucun logiciel ou programme n'est disponible).

Enfin, Goldman et Hannan (Goldman & Hannan, 2001) proposent un schéma permettant une surveillance en continue (c'est-à-dire après inclusion de chaque nouveau patient) de la toxicité d'un traitement expérimental dont l'efficacité est évaluée de façon non continue au cours d'un essai simple bras. Le nombre de sujets nécessaire est déterminé par l'essai mesurant l'efficacité et les règles d'arrêt pour toxicité sont fixées à partir de cet effectif, des taux de toxicité sous les hypothèses nulle et alternative et de l'erreur de type I fixés sur la toxicité. Un programme est disponible auprès des auteurs. Le schéma prévoit de ne pas s'arrêter après une seule toxicité mais uniquement si deux toxicités sont observées sur un même patient, trois toxicités parmi deux patients,…jusqu'à neuf toxicités parmi huit patients. Cette approche ne

tient pas compte de l'effet de cette surveillance de la toxicité sur les hypothèses d'activité. Les auteurs notent également que ce schéma ne convient qu'aux petits essais d'environ 20-30 patients.

Ivanova et al (Ivanova, 2005) proposent également un schéma d'évaluation continue de la toxicité parallèlement à un schéma en une étape conduit pour évaluer l'activité. La toxicité est évaluée après la participation de chaque patient parfois même après l'inclusion du patient suivant si les délais d'inclusion sont plus courts que les délais d'observation de la toxicité. Le nombre de sujets nécessaire est calculé à partir du critère d'efficacité et sert à déterminer les bornes d'arrêt pour la toxicité. Les critères d'évaluation de la toxicité et de l'efficacité sont binaires et, contrairement au schéma précédent, la corrélation entre les deux critères doit être définie lors de la planification de l'étude car l'évaluation de la toxicité peut avoir un impact sur l'estimation de la réponse. Les auteurs soulignent cependant que sauf si la toxicité et l'efficacité sont très corrélées, cet impact reste minime. Cette méthode utilise les bornes d'arrêt de Pocock ou O'Brian et Fleming. Un logiciel est disponible sur : http ://www.bios.unc.edu/distrib/gee/crossing/cp3/.

En 2012, Ray et Rai (Ray & Rai, 2012) ont proposé une extension de la méthode d'Ivanova en se basant sur un plan de Simon multi-étapes (Simon, 1989). Les deux critères sont considérés comme indépendants mais les auteurs soulignent que la dépendance entre l'efficacité et la tolérance a en réalité un impact sur les effectifs moyens et la probabilité d'arrêt précoce.

Au cours de ces dernières années, plusieurs méthodologies de phase II permettant de prendre en compte à la fois les données de tolérance et d'efficacité ont été développées, selon des approches fréquentistes comme bayésiennes. Certaines proposent une analyse séquentielle des événements toxiques, au même rythme que les données sur la réponse. Une analyse continue, après inclusion de chaque patient, ou du moins plus fréquente est sans doute mieux adaptée aux événements indésirables graves. Le fait de considérer deux critères suppose d'emblée que l'on puisse recueillir

l'information pour les deux événements. La toxicité à long terme peut donc difficilement être étudiée. Il faut également penser au problème de la censure d'un événement sur le second. En effet, la survenue d'une toxicité faisant interrompre le traitement par exemple (voire entraînant le décès), peut conduire à classer le patient en échec pour la réponse. De la même façon, une diminution des doses du fait de la survenue d'une toxicité peut entraîner une perte d'efficacité du traitement. Finalement, cela peut modifier quantitativement voire qualitativement l'association observée entre les deux critères et donc laisser à penser que l'hypothèse faite sur la relation entre réponse et toxicité lors de la rédaction du protocole était erronée. Toutefois, la constatation d'un tel phénomène peut aussi devenir un atout. La prise en compte de la tolérance à ce stade de l'évaluation thérapeutique permet de favoriser le développement des essais de désescalade de doses. Ainsi, si l'on observe dans un essai beaucoup d'arrêts ou d'adaptations des doses de traitement, on peut proposer un autre essai à une dose moindre, mieux tolérée qui se révèlerait être peut-être plus efficace car moins limitante.

METHODOLOGIES PERMETTANT D'EVALUER L'ACTIVITE PAR STRATE

Pour affirmer le bénéfice spécifique d'une thérapeutique sur une sous-population, l'une des options offertes aux investigateurs est de conduire un essai pour chaque sous-groupe de population, entraînant automatiquement une augmentation importante du temps nécessaire à l'évaluation, du coût et surtout du nombre de personnes participant à l'expérimentation. Une question éthique est alors soulevée puisqu'il est indispensable de minimiser le nombre de patients ne bénéficiant pas de la thérapeutique et donc exposés inutilement à une thérapeutique expérimentale.

L'autre option est de tenir compte de l'hétérogénéité de la population dès la phase II et dans le schéma d'étude mis en place, pour être en capacité d'identifier éventuellement une sous-population sensible au sein d'un groupe

hétérogène de patients qui, étudiés dans son ensemble, sembleraient ne pas pouvoir bénéficier de la nouvelle thérapeutique.

Les principales méthodes développées ces dernières années pour tenir compte de l'hétérogénéité des patients sont citées dans le Tableau 4.

Parmi les premières publiées, on peut citer celle proposée par London et Chang (London & Chang, 2005) qui permet d'évaluer le taux de réponse au niveau de différentes strates. Les critères de jugement sont binaires et le schéma est à un seul bras et à une ou deux étapes. Deux approches sont proposées : l'une dans laquelle la proportion de patients dans chaque strate est connue (approche non conditionnelle), et l'autre dans laquelle cette proportion est inconnue (approche conditionnelle). Un algorithme de simulation est donné pour calculer le nombre de sujets nécessaire et les règles d'arrêt de l'approche non conditionnelle. Des calculs exacts sont utilisés pour l'approche conditionnelle. Des détails sont donnés dans l'article mais les deux approches nécessitent de la programmation pour pouvoir être utilisées.
Le test statistique de l'approche conditionnelle est basé sur le nombre total de réponses observées conditionnellement aux effectifs observés de chaque strate. Cette approche est celle recommandée par les auteurs puisqu'il n'est pas nécessaire de connaître la proportion exacte de chaque strate au début de l'étude et semble d'après les auteurs robuste aux hypothèses faites sur les proportions d'effectifs attendues. Si les cliniciens ont des difficultés à déterminer ces proportions, plusieurs hypothèses peuvent être faites et le plus large effectif total conservé. Pour l'approche non conditionnelle à deux étapes que les auteurs proposent également, les cliniciens doivent spécifier la proportion des erreurs de type I et II globales dépensée à la première étape. Un arrêt précoce pour inefficacité est possible. Qu'elle soit conditionnelle ou non conditionnelle, la méthodologie proposée permet effectivement de tenir compte de l'hétérogénéité de la population mais ne permet pas d'identifier une sous-population plus sensible au traitement s'il y a lieu.

Une extension de l'approche conditionnelle a été proposée en 2009 par Sposto et Gaynon (Sposto & Gaynon, 2009) permettant de tenir compte des prévalences observées des différentes strates lors de la première étape, tout en contrôlant les risques d'erreur. Il n'y a pas d'ajustement des effectifs car les auteurs anticipent dès l'élaboration du protocole toutes les règles d'arrêt possibles en fonction des prévalences possibles de chaque strate.
Deux autres extensions de l'approche conditionnelle ont été publiées en 2012. L'une par Jung et al (Jung, 2012) permettant un arrêt précoce pour inefficacité suivant la méthode proposée par Simon (Simon, 1989) mais avec des bornes d'arrêt modifiées en fonction des prévalences observées des différentes strates, l'autre a été proposée par Chang et al (Chang, 2012) et permet un ajustement des effectifs de chaque strate à la deuxième étape et ainsi une amélioration de la puissance.

Jones et Holmgren (Jones & Holmgren, 2007) proposent un schéma à deux étapes tenant compte de la positivité à un biomarqueur, disponible au début de l'étude et dont la prévalence est connue. Ce schéma est une extension de la méthode de Simon à deux étapes. Dans ce schéma les patients positifs et les patients négatifs pour le biomarqueur sont clairement distingués dès le début de l'étude alors que Pusztai (Pustzai 2007, décrit ci-dessous) considère dans un premier temps l'ensemble des patients puis le sous-groupe des positifs uniquement si cela est approprié. Ce schéma étant comme précédemment adapté de la méthode de Simon, seul un arrêt précoce pour inefficacité est possible.

Pustzai (Pustzai 2007) propose un schéma à deux étapes et un seul bras qui permet d'évaluer la réponse à un traitement expérimental en fonction de la positivité à un biomarqueur. L'objectif est de déterminer si la nouvelle thérapeutique a une efficacité pour un ensemble de patients non sélectionnés ou si le niveau d'efficacité intéressant n'est atteint que pour le sous-groupe de patients positifs pour le biomarqueur. Ce biomarqueur doit être spécifié avant le début de l'étude. Plusieurs marqueurs peuvent être étudiés. Le calcul

des effectifs et des bornes d'arrêt est basé sur la méthode de Simon (Simon, 1989).

L'essai se déroule comme suit :

1) Recrutement classique selon un schéma à deux étapes de Simon, sans tenir compte du biomarqueur.
2) Si à la fin de l'étape 1, suffisamment de réponses sont observées pour aller en étape 2, poursuite de l'inclusion de façon classique sans tenir compte du biomarqueur et conclure à l'étape 2 selon la méthode classique pour l'ensemble de la population.
3) Si à la fin de l'étape 1, le nombre total de réponses est trop faible, mise en œuvre d'un nouvel essai de phase II en deux étapes uniquement avec des patients positifs pour le biomarqueur.
4) Considérant 3) si à la fin de l'étape additionnelle, le nombre total de réponses est toujours trop faible, arrêt de l'essai pour inefficacité.
5) Considérant 3) si à la fin de l'étape 1 du nouvel essai, le nombre total de réponses est suffisamment élevé pour justifier de la poursuite de l'étude, mettre en œuvre la deuxième étape.
6) Considérant 5), Conclure sur l'efficacité ou non de la nouvelle thérapeutique pour le sous-groupe des patients positifs au biomarqueur.

Si plusieurs marqueurs sont étudiés, alors ce sont autant de nouveaux essais de phase II qui sont conduits. La méthodologie étant basée sur un plan de Simon, l'arrêt pour efficacité dès la première étape n'est pas possible.

Le Blanc (Le Blanc, 2009) propose un schéma séquentiel qui permet l'inclusion d'une large population quand la population cible spécifique de la nouvelle thérapeutique est incertaine, de telle sorte que les hypothèses testées et les analyses conduites portent sur plusieurs sous-groupes de patients. Les critères de jugement sont binaires mais la méthodologie peut être adaptée aux critères de survie. L'évaluation de l'efficacité se fait dans chaque strate après inclusion des premiers patients. Un arrêt précoce pour inefficacité est possible dans seulement l'une des strates ou pour l'ensemble de la population si le test global est négatif.

Thall (Thall et al, 2003) puis Wathen plus récemment (Wathen, 2008) ont proposé des approches bayésiennes. Wathen et al proposent un schéma selon une approche bayésienne pour des essais à un seul bras, séquentiels ou avec évaluation continue de critères binaires ou de survie, et tenant compte de l'hétérogénéité des patients. Le schéma proposé généralise les approches de Thall et Simon (Thall & Simon, 1994) et de Thall et al (Thall, 2005) et utilise la régression linéaire avec un terme d'interaction entre le traitement et la variable sous-groupe. Les effets du traitement peuvent donc varier entre les sous-groupes. Un arrêt précoce pour inefficacité est possible. Des simulations sont nécessaires pour valider le schéma au regard des taux de faux-négatifs et de faux-positifs et un nombre maximal de sujets nécessaire doit être spécifié. Aucun logiciel n'est disponible.

Tableau 4. ***Liste des méthodes permettant la prise en compte l'existence de strates dans les essais de phase II***

	Auteurs
1 étape	A'Hern (2004)
	London & Chang (2005)
2 étapes	A'Hern (2004)
	London & Chang (2005)
	Putzai (2007)
	Jones & Holmgren (2007)
	Sposto (2009)
	Jung (2012)
	Chang (2012)
Multi-étapes	Thall et al (2003)
	Wathen et al (2008)
	Le Blanc et al (2009)
Surveillance en continue	Thall et al (2003)
	Wathen et al (2008)

Les méthodologies développées ces dernières années et permettant d'évaluer l'activité par strate présentent certains défauts : elles ne permettent pas d'arrêter précocement l'essai pour efficacité alors que cette situation permet d'éviter à des patients non sensibles d'être inutilement exposés à des traitements potentiellement toxiques pour eux ; elles ne permettent pas systématiquement la sélection d'une sous-population sensible mais ne font que tenir compte de l'hétérogénéité de la population pour l'estimation de la réponse globale. Certaines d'entre elles ont par contre l'avantage de permettre la prise en compte de plusieurs biomarqueurs et donc de plusieurs strates sachant que dans ces circonstances, seule une strate peut être retenue à la fin de l'essai.

Ainsi, les chapitres suivants présentent la méthodologie développée au cours de ce travail et son application en pratique clinique. La méthodologie proposée est une extension de la méthode de Fleming en deux étapes. L'effet du traitement est donc mesuré par un critère binaire (succès ou échec) et la méthode permet l'arrêt précoce de l'essai pour efficacité. Elle considère deux strates, définie sur la présence ou non d'une caractéristique particulière du patient. Elle permet de déterminer à la fin de l'essai si le développement doit être arrêté ou poursuivi pour toute la population ou seulement une partie. Il ne s'agit donc pas seulement de prendre en compte l'hétérogénéité de la population dans l'évaluation de l'effet du médicament mais bien d'une recherche d'amélioration de la définition de la population cible de la nouvelle thérapeutique.

CHAPITRE 2. NOUVELLE METHODOLOGIE DE PHASE II PERMETTANT DE PRENDRE EN COMPTE L'HETEROGENEÏTE DE LA POPULATION PARTICIPANT A LA RECHERCHE

Dans cette partie, nous présentons une nouvelle approche publiée dans la revue Statistics in Medicine (Tournoux-Facon, 2011a, annexe 2).

Cette nouvelle méthode pour les essais de phase II permet de déterminer en cas d'inefficacité apparente sur l'ensemble de la population si un sous-groupe ne serait pas sensible au traitement afin de pouvoir continuer le développement du médicament pour ce sous-groupe particulier et ne pas conclure à tort à l'inefficacité de la nouvelle thérapeutique pour l'ensemble de la population.

A l'inverse, elle permet également en cas d'efficacité apparente sur l'ensemble de la population, de déterminer si ce n'est pas seulement un sous-groupe qui serait sensible au traitement afin de pouvoir arrêter le développement du médicament pour l'autre sous-groupe et ne pas l'exposer plus longtemps à une toxicité éventuelle de la nouvelle thérapeutique.

Pour cela, nous avons développé une méthode qui est, contrairement à tous les autres schémas déjà publiés, une extension du plan de Fleming (Fleming, 1982) que nous rappelons dans un premier temps.

RAPPELS SUR LE PLAN DE FLEMING CLASSIQUE A 2 ETAPES

Thomas R. Fleming a publié en 1982 une méthode proposant des plans en une ou plusieurs étapes permettant l'arrêt précoce de l'étude dans le cas où les premiers résultats montrent de façon indiscutable l'efficacité ou l'inefficacité du traitement étudié.

NOTATIONS ET HYPOTHESES

a) Effectifs

La méthode considère que les patients intègrent l'essai séquentiellement suivant deux étapes $s = 1\ ou\ 2$.
Les effectifs n_s aux première et deuxième étapes sont notés n_1 et n_2.
Les effectifs cumulés sont notés N_1 et N_2.
A la première étape N_1=n_1,
A la deuxième étape, N_2=$n_1 + n_2$.
Les effectifs peuvent ne pas être identiques entre les deux étapes.

b) Critère de jugement

La réponse au traitement est binaire, succès ou échec, et suit une loi de Bernoulli.
r_s est défini comme étant le nombre de succès observés à l'issue de chaque étape s sur n_s sujets et suit une loi Binomiale.
Le nombre cumulé de réponses observées à chaque étape est noté R_s .
A la première étape R_1=r_1,
A la deuxième étape, R_2=$r_1 + r_2$.

c) Hypothèses nulle et alternative

Doivent également être définis les paramètres π_1 et π_0 représentant respectivement les seuils d'efficacité et d'inefficacité.

Le paramètre π_0 représente la probabilité d'efficacité minimale et π_1 la probabilité d'efficacité que l'on souhaite absolument mettre en évidence si elle existe (efficacité optimale).

La probabilité réelle de succès est définie par le paramètre π.

Les 2 hypothèses à considérer sont :

$H_0 : \pi \leq \pi_0$ (H_0 : hypothèse nulle d'inefficacité)

$H_1 : \pi > \pi_0$ (H_1 : hypothèse alternative d'efficacité).

Ainsi :

- Si $\pi \leq \pi_0$, la molécule étudiée sera considérée comme insuffisamment efficace.
- Si $\pi > \pi_0$, la molécule étudiée sera considérée comme suffisamment efficace pour entreprendre des essais de phase III.

d) Risques d'erreur

Les risques de première espèce α et de deuxième espèce β sont définis en tenant compte des conséquences de retenir à tort une molécule inefficace ou au contraire de rejeter à tort une molécule efficace.

A noter que le risque de deuxième espèce β est calculé sous l'hypothèse alternative particulière $\pi = \pi_1$.

Le Tableau 5 ci-dessous récapitule les notations utilisées pour décrire un plan de Fleming classique à deux étapes.

Tableau 5. Notations utilisées pour décrire un plan de Fleming classique à deux étapes

Paramètres	Notation
Efficacité minimale	π_0
Efficacité optimale	π_1
Efficacité réelle	π
Hypothèse nulle	$H_0 : \pi \leq \pi_0$
Hypothèse alternative	$H_1 : \pi > \pi_0$
Erreur de type I	α
Erreur de type II	β
Etape	s
Effectif par étape	n_s
Effectif cumulé à l'étape 1	N_1
Effectif cumulé à l'étape 2	$N_2 = n_1 + n_2$
Succès par étape	r_s
Succès cumulé à l'étape 1	$R_1=r_1$
Succès cumulé à l'étape 2	$R_2=r_1 + r_2$

PRINCIPES GENERAUX

A chaque étape, une décision est prise en fonction des résultats observés (Tableau 6).

La méthode consiste à calculer 4 bornes d'arrêt (a_1, b_1, a_2, b_2) qui permettent de définir les règles d'arrêt suivantes. Le plan de Fleming est un plan fermé (une conclusion est obligatoire à la dernière étape).

A la dernière étape, $a_2=b_2$-1.

Tableau 6. Règles d'arrêt lors d'un Fleming classique à 2 étapes

		$R_1 \leq a_1$	$R_1 \in]\, a_1 - b_1\, [$	$R_1 \geq b_1$
Etape 1		Arrêt pour Inefficacité	Poursuite en étape 2	Poursuite en phase III
Etape 2	$R_2 < b_2$	-	Arrêt pour Inefficacité	-
	$R_2 \geq b_2$	-	Poursuite en phase III	-

On définit la variable aléatoire Ω_1 qui prend les valeurs suivantes :

- $\Omega_1 = -1$ quand $R_1 \leq a_1$
- $\Omega_1 = 0$ quand $R_1 \in]\, a_1 - b_1\, [$
- $\Omega_1 = 1$ quand $R_1 \geq b_1$.

On définit également la variable aléatoire Ω_2 qui prend les valeurs suivantes :

- $\Omega_2 = -1$ quand $R_2 < b_2$
- $\Omega_2 = 1$ quand $R_2 \geq b_2$.

A la première étape, on inclut n_1 sujets.

A la fin de la première étape, l'efficacité du traitement est considérée comme insuffisante si le nombre cumulé de succès est inférieur ou égal à la valeur a_1 ($\Omega_1 = -1$). Il est conclu à une efficacité insuffisante de la thérapeutique pour être développée en phase III.
Par contre, si le nombre cumulé de succès est supérieur ou égal à une valeur b_1 ($\Omega_1 = 1$) l'hypothèse nulle est rejetée et la molécule est considérée comme suffisamment efficace pour être développée en phase III. Si le nombre de succès est compris entre a_1 et b_1 ($\Omega_1 = 0$), n_2 sujets supplémentaires sont inclus dans l'étape 2.

A la deuxième étape, on conclut à l'inefficacité de la thérapeutique si $\Omega_2 = -1$ ou à son efficacité si $\Omega_2 = 1$.

CALCULS

a) Bornes d'arrêt

Pour le calcul des bornes d'arrêt, on se donne un couple (n_1, n_2).

Pour chaque étape, les limites de décision a_s et b_s sont calculées selon les formules suivantes :

$$a_s = \begin{cases} 0 & a'_s < 0 \\ Entier\ Inférieur(a'_s) & a'_s \geq 0 \end{cases}$$

$$b_s = Entier\ Supérieur\left(N_s\pi_0 + z_{1-\alpha}\sqrt{N_2\pi_0(1-\pi_0)}\right)$$

Avec :

$$a'_s = \left(N_s \times p - z_{1-\alpha}\sqrt{N_2 p(1-p)}\right)$$

$$p = \frac{\left[\sqrt{N_2\pi_0} + z_{1-\alpha}\sqrt{1-\pi_0}\right]^2}{N_2 + z_{1-\alpha}^2}.$$

Notons que β n'intervient pas dans les formules.

b) Effectifs

Pour l'effectif total calculé initialement, on obtient deux valeurs de a_s *et* b_s. On calcule alors la puissance obtenue avec cette configuration. Si celle-ci est inférieure à $1-\beta$, on augmente le couple (n_1, n_2) et on recalcule les bornes pour tendre le plus possible vers $1-\beta$ et inversement si celle-ci est supérieure à $1-\beta$: on diminue le couple (n_1, n_2).

On itère ce processus jusqu'à obtenir une puissance calculée supérieure ou égale à $1-\beta$, avec l'effectif total le plus petit possible.

On peut se donner une estimation de l'effectif total à choisir pour initier l'itération à partir de la formule suivante (Fleming à une étape) :

$$n = \left[\frac{z_{1-\alpha}\sqrt{\pi_0(1-\pi_0)} + z_{1-\beta}\sqrt{\pi_1(1-\pi_1)}}{(\pi_1 - \pi_0)}\right]^2.$$

Par habitude, les effectifs à chaque étape sont égaux. L'augmentation lors de l'itération consiste alors à ajouter ou soustraire 2 sujets (1 par étape). Mais il est également possible de déséquilibrer le nombre de sujets par étape et donc de ne faire varier lors des itérations que l'un ou l'autre des effectifs.

PRESENTATION DE LA NOUVELLE METHODE

La nouvelle méthode que nous proposons est une extension du plan de Fleming à deux étapes. La population des patients inclus est scindée en deux sous-populations ou strates sur une caractéristique clinique ou biologique (biomarqueur) connue pour chaque patient avant le début de l'essai et qui d'après les données de la littérature ou en se basant sur les mécanismes d'actions physio-pathologiques du traitement, pourrait influencer le taux de réponses observé à la fin de l'essai.

Nous appellerons cette nouvelle méthode « Méthode A » par la suite en opposition à la méthode de Fleming classique dite « Méthode H » pour Hétérogène, puisque ne tenant pas compte de l'existence de deux sous-populations.

L'originalité de la méthode A par rapport à la Méthode H est la prise en compte lors de la planification de l'essai d'une caractéristique particulière dont on pense qu'elle pourrait influencer le taux de réponse observés. L'enjeu est donc de déterminer à la fin de la première ou éventuellement à la fin de la deuxième étape si les réponses observées entre les deux sous-populations sont hétérogènes au point de justifier l'arrêt de l'essai pour l'une des deux sous-populations (pour efficacité ou inefficacité).

Il n'est pas question de ne sélectionner que la sous-population qui répond le mieux, mais d'arrêter l'essai au plus vite pour une sous-population qui ne bénéficie pas du tout de la nouvelle thérapeutique ou à l'inverse poursuivre l'évaluation pour la seule sous-population dont on est certain qu'elle en bénéficie afin d'accélérer le développement de la thérapeutique.

NOTATIONS ET HYPOTHESES

a) Effectifs

La méthode considère deux sous populations $i = 1\ ou\ 2$ prédéfinies.

Les patients intègrent l'essai séquentiellement suivant deux étapes $s = 1\ ou\ 2$.

Les effectifs aux première et deuxième étapes sont notés n_{i1} et n_{i2} pour chaque sous-population i.

Sont donc inclus :

- Etape 1, $n_{11} + n_{21} = n_{\cdot 1}$ patients
- Etape 2, $n_{12} + n_{22} = n_{\cdot 2}$ patients.

Le nombre cumulé de patients inclus depuis le début de l'essai au sein d'une sous-population i et à l'étape s est noté N_{is} :

- Etape 1 : $N_{i1} = n_{i1}$
- Etape 2 : $N_{i2} = n_{i1} + n_{i2}$.

Le nombre cumulé de patients inclus à chaque l'étape s est noté $N_{\cdot s}$:

- Etape 1 : $N_{\cdot 1} = n_{11} + n_{21} = n_{\cdot 1}$
- Etape 2 : $N_{\cdot 2} = n_{11} + n_{21} + n_{12} + n_{22} = n_{\cdot 1} + n_{\cdot 2}$.

Le nombre cumulé de patients inclus en fin d'essai dans chaque sous-population i est noté $N_{i\cdot}$:

- Sous-population 1 : $N_{1\cdot} = n_{11} + n_{12}$
- Sous-population 2 : $N_{2\cdot} = n_{21} + n_{22}$.

Nous faisons l'hypothèse que le ratio entre les deux sous-populations peut être différent de 1 mais qu'il est constant entre les deux étapes et connu *a priori* : $n_{2s} = \omega \times n_{1s}$ (1)

b) Critère de jugement

Le critère de jugement principal mesurant l'efficacité est un critère binaire, succès ou échec. Pour le j^{th} patient de la sous-population i à l'étape s, ce critère est noté X_{isj}, avec X_{isj}=1 en cas de réponse (succès) et 0 en cas de non réponse (échec).

Le nombre de réponses observées depuis le début de l'essai au sein d'une sous-population i et lors de l'étape s est noté :

$$r_{is} = \sum_{j=1}^{n_{is}} X_{isj}.$$

Le nombre cumulé de réponses observées depuis le début de l'essai au sein d'une sous-population i et à l'étape s est noté R_{is} :

- Etape 1 : $R_{i1}= r_{i1}$
- Etape 2 : $R_{i2}= r_{i1} + r_{i2}$.

Le nombre cumulé de réponses observées à chaque l'étape s est noté $R_{.s}$:

- Etape 1 : $R_{.1}= r_{11} + r_{21}= r_{.1}$
- Etape 2 : $R_{.2}= r_{11} + r_{21} + r_{12} + r_{22} = r_{.1} + r_{.2}$.

Le nombre cumulé en fin d'essai des réponses observées dans chaque sous-population i est noté $R_{i.}$:

- Sous-population 1 : $R_{1.}= r_{11} + r_{12} = R_{12}$
- Sous-population 2 : $R_{2.}= r_{21} + r_{22} = R_{22}$.

c) Hypothèses nulles et alternatives

Le paramètre π_{0i} représente la probabilité d'efficacité minimale dans la sous-population i et π_{1i} la probabilité d'efficacité optimale (ou probabilité d'efficacité que l'on souhaite absolument mettre en évidence si elle existe) dans la sous-population i, avec $\Delta_i = \pi_{1i} - \pi_{0i}$.

π_{01} peut être différente de π_{02} et Δ_1 peut être différent de Δ_2 également.

π_1 et π_2 sont les probabilités réelles de succès des sous-populations 1 et 2.
Dans ce travail, nous définissions les hypothèses nulle et alternative suivantes :

- Hypothèse nulle : $\pi_1 \leq \pi_{01}$ et $\pi_2 \leq \pi_{02}$ (H_0 : Aucune des deux sous-populations n'est sensible à la nouvelle thérapeutique)
- Hypothèse alternative : $\pi_1 > \pi_{01}$ ou $\pi_2 > \pi_{02}$ (H_1 : Au moins l'une des deux sous-populations est sensible à la nouvelle thérapeutique)

Au sein des hypothèses nulles et alternatives, on peut définir une hypothèse nulle H_{00} et une hypothèse alternative particulière H_{11} qui seront utilisées pour le calcul des risques d'erreur :

H_{00} : $\pi_1 = \pi_{01}$ et $\pi_2 = \pi_{02}$
H_{11} : $\pi_1 = \pi_{11}$ et $\pi_2 = \pi_{12}$.

Par ailleurs nous utiliserons deux autres hypothèses alternatives particulières :

H_{01} : $\pi_1 = \pi_{01}$ et $\pi_2 = \pi_{12}$
H_{10} : $\pi_1 = \pi_{11}$ et $\pi_2 = \pi_{02}$.

d) Risques d'erreur

La décision de rejeter ou non l'hypothèse nulle H_0 est notée Φ, variable binaire :

- Φ=1 correspond au rejet de l'hypothèse nulle et consiste donc à déclarer la nouvelle thérapeutique efficace pour au moins l'une des deux sous-populations,
- Φ=0 correspond au non-rejet de l'hypothèse nulle et consiste donc à ne pas considérer la nouvelle thérapeutique efficace sur aucune des deux sous-populations (inefficace pour la sous-population 1 et inefficace pour la sous-population 2).

La décision Φ dépend des données observées au sein des deux sous-populations au cours des deux étapes.

Pour la méthode A, l'erreur de type I consiste à conclure à la fin de l'essai à l'efficacité de la thérapeutique dans au moins une des deux sous-populations alors qu'aucune des deux sous-populations n'est sensible à la nouvelle thérapeutique (on se place donc sous l'hypothèse nulle limite H_{00}).

On définit la puissance comme la probabilité de déclarer la nouvelle thérapeutique efficace pour au moins une sous-population sous H_{11}.

Le schéma doit donc satisfaire les contraintes sur les risques d'erreurs de type I et II notées α et β qui s'écrivent de la façon suivante :

$$P\{\Phi = 1 \mid \pi_1 = \pi_{01} \text{ et } \pi_2 = \pi_{02}\} \leq \alpha \text{ et } P\{\Phi = 0 \mid \pi_1 = \pi_{11} \text{ et } \pi_2 = \pi_{12}\} \leq \beta.$$

Dans le cadre de la Méthode A, on introduit également le risque γ associé à la probabilité, d'identifier à chaque étape une hétérogénéité de réponses entre les deux sous-populations sous l'hypothèse nulle limite H_{00}.

Le Tableau 7 ci-dessous récapitule les notations utilisées pour décrire le nouveau plan (Méthode A) comparativement au plan de Fleming classique à deux étapes (Méthode H).

Tableau 7. Notations utilisées pour décrire le nouveau plan comparativement au plan de Fleming classique à deux étapes

	Méthode H	Méthode A
Sous-population	-	i
Paramètres		
Efficacité minimale	π_0	π_{0i}
Efficacité optimale	π_1	π_{1i}
Efficacité réelle	π	π_i
Hypothèse nulle	$H_0 : \pi \leq \pi_0$	$H_0 : \pi_1 \leq \pi_{01}$ et $\pi_2 \leq \pi_{02}$
Hypothèse alternative	$H_1 : \pi > \pi_0$	$H_1 : \pi_1 > \pi_{01}$ ou $\pi_2 > \pi_{02}$
Erreur de type I	α	α
Erreur de type II	β	β
Identification d'une hétérogénéité sous H_{00}	-	γ
Etape	s	s
Effectif par étape	n_s	$n_{\cdot s}$
Effectif cumulé à l'étape 1	N_1	$N_{\cdot 1} = n_{11} + n_{21}$
Effectif cumulé à l'étape 2	$N_2 = n_1 + n_2$	$N_{\cdot 2} = (n_{11} + n_{21}) + (n_{12} + n_{22})$
Succès par étape	r_s	r_{is}
Succès cumulés à l'étape 1	$R_1 = r_1$	$R_{\cdot 1} = r_{11} + r_{21}$
Succès cumulés à l'étape 2	$R_2 = r_1 + r_2$	$R_{\cdot 2} = r_{11} + r_{21} + r_{12} + r_{22}$

PRINCIPES GENERAUX

Les bornes d'arrêt et les effectifs sont calculés comme dans un plan de Fleming classique, c'est-à-dire sans tenir compte de l'existence des strates.

La méthode A est comme le plan de Fleming un plan fermé.

A chaque étape, on a la possibilité de déterminer une hétérogénéité entre les réponses des deux sous-populations, notée Ψ_s :

1) Ψ_s=0 : Aucune hétérogénéité n'est détectée entre les deux sous-populations
2) Ψ_s=1 : Identification d'une hétérogénéité entre les réponses en faveur de la sous-population 1
3) Ψ_s=2 : Identification d'une hétérogénéité entre les réponses en faveur de la sous-population 2

A l'étape 1, on inclut $N_{\cdot 1}$ sujets.

Le nombre de réponses observées ($R_{.1}$), couplée à l'existence d'une hétérogénéité en faveur de la sous-population i ou non, permet de définir 9 règles de décision à l'étape 1 contre 3 avec le plan de Fleming classique (Tableau 8) :

On note :

- I_i=Inefficacité pour la sous-population i
- E_i=Efficacité pour la sous-population i
- C_i=Poursuite des investigations en étape 2

Tableau 8. Règles d'arrêt à la fin de l'étape 1

	$\Omega_1 = -1$	$\Omega_1 = 0$	$\Omega_1 = 1$
Ψ_1**= 1**	$C_1 I_2$	$C_1 I_2$	$E_1 I_2$
Ψ_1**= 0**	$I_1 I_2$	$C_1 C_2$	$E_1 E_2$
Ψ_1**= 2**	$I_1 C_2$	$I_1 C_2$	$I_1 E_2$

I_i=Inefficacité pour la sous-population i ; E_i=Efficacité pour la sous-population i ; C_i=Poursuite des investigations en étape 2 avec la sous-population i.

- Si Ψ_1=1 et $\Omega_1 < 1$, l'évaluation de la nouvelle thérapeutique est poursuivie en étape 2 avec uniquement la sous-population 1 et s'arrête en étape 1 pour la sous-population 2 (situation $C_1 I_2$). Dans un modèle de Fleming classique l'évaluation aurait soit été arrêtée ($\Omega_1 = -1$) soit poursuivie en étape 2 ($\Omega_1 = 0$) pour l'ensemble de la population, alors que dans ce contexte on peut penser que la sous-population 1 pourrait bénéficier de la thérapeutique et que l'inefficacité apparente est en réalité quasi-exclusivement liée aux échecs au sein de la sous-population 2.

- Si Ψ_1=1 et $\Omega_1 = 1$, l'évaluation de la nouvelle thérapeutique est poursuivie en phase III avec uniquement la sous-population 1 (situation $E_1 I_2$). Elle s'arrête en étape 1 pour la sous-population 2, car la détermination d'une hétérogénéité est en défaveur de la sous-population 2. Le franchissement de la borne b_1 est donc uniquement lié aux excellents résultats observés dans la sous-population 1. Il n'y a donc aucun intérêt à continuer l'évaluation thérapeutique sur

l'ensemble de la population comme on serait amené à le faire avec le plan de Fleming classique au risque d'exposer inutilement la sous-population 2 à une thérapeutique potentiellement toxique pour elle et de diluer l'effet traitement en deuxième étape ou en phase III si la prévalence de la sous-population 2 est élevée.

- Si Ψ_1=2 et $\Omega_1 < 1$, l'évaluation de la nouvelle thérapeutique est poursuivie en étape 2 avec uniquement la sous-population 2 et s'arrête en étape 1 pour la sous-population 1 (situation I_1C_2). Dans un modèle de Fleming classique l'évaluation aurait soit été arrêtée ($\Omega_1 = -1$) soit poursuivie en étape 2 ($\Omega_1 = 0$) pour l'ensemble de la population, alors que dans ce contexte on peut penser que la sous-population 2 pourrait bénéficier de la thérapeutique est que l'inefficacité apparente est en réalité quasi-exclusivement liée aux échecs au sein de la sous-population 1.

- Si Ψ_1=2 et $\Omega_1 = 1$, l'évaluation de la nouvelle thérapeutique est poursuivie en phase III avec uniquement la sous-population 2 (situation I_1E_2). Elle s'arrête en étape 1 pour la sous-population 1, car la détermination d'une hétérogénéité est en défaveur de la sous-population 1. Il n'y a donc aucun intérêt à continuer l'évaluation thérapeutique sur l'ensemble de la population comme on serait amené à le faire avec le plan de Fleming classique au risque d'exposer inutilement la sous-population 1 à une thérapeutique potentiellement toxique pour elle et de diluer l'effet traitement en deuxième étape ou en phase III si la prévalence de la sous-population 1 est élevée.

- Si Ψ_1=0 cela signifie qu'aucune hétérogénéité de réponse n'a été détectée entre les deux sous-populations. Les décisions sont les mêmes que lors d'un plan de Fleming classique :
    - Arrêt de l'ensemble de la population si $\Omega_1 = -1$ (situation I_1I_2)
    - Poursuite en étape 2 de l'ensemble de la population si $\Omega_1 = 0$ (situation C_1C_2)

- Poursuite en phase III de l'ensemble de la population si $\Omega_1 = 1$ (situation E_1E_2).

A la fin de l'étape 1, on peut donc :

- arrêter les deux sous-populations pour inefficacité ou efficacité
- continuer avec une seule sous-population
- continuer avec les deux sous-populations

Si l'ensemble de la population poursuit l'essai en étape 2, on inclut $n_{12} + n_{22}$ sujets supplémentaires.

Le nombre de réponses observées ($R_{.2}$), couplée à l'existence d'une hétérogénéité en faveur de la sous-population i ou non, permet de définir 4 règles de décision à l'étape 2 contre 2 avec le plan de Fleming classique (Tableau 9) :

Tableau 9. Règles d'arrêt à la fin de l'étape 2

	$\Omega_2 = -1$	$\Omega_2 = 1$
$\Psi_2 = 1$	I_1I_2	E_1I_2
$\Psi_2 = 0$	I_1I_2	E_1E_2
$\Psi_2 = 2$	I_1I_2	I_1E_2

I_i=Inefficacité pour la sous-population i ; E_i=Efficacité pour la sous-population i

Si une seule sous-population continue de participer à l'essai en étape 2 (situations C_1I_2, et I_1C_2), le nombre de patients recrutés pour l'étape 2 et la borne b_2 doivent être adaptés pour contrôler les risques α et β.

Cette adaptation est anticipée dès l'élaboration du protocole de telle sorte que l'investigateur connaît avant même le début de l'étude le nombre maximal de patients susceptibles d'être inclus au total.

Ce sont deux plans de Fleming, indépendants qui sont élaborés :

- plan de Fleming indépendant (F1) au cas où seule la sous-population 1 continuerait en étape 2 auquel on associe Ω_2F1 et n_2F1 (nombre de sujets supplémentaires de la sous-population 1 inclus à l'étape 2)

- plan de Fleming indépendant (F2) au cas où seule la sous-population 2 continuerait en étape 2 auquel on associe Ω_2F2 et n_2F2 (nombre de sujets supplémentaires de la sous-population 2 inclus à l'étape 2)

Les règles de décision dans le cas où on continue avec la sous-population 1 deviennent (Tableau 10) :

Tableau 10. ***Règles d'arrêt à la fin de l'étape 2 lorsque seule la sous-population 1 continue après l'étape 1,***

Ω_2F1 = −1	Ω_2F1 = 1
$I_1 I_2$	$E_1 I_2$

I_i=Inefficacité pour la sous-population i ; E_i=Efficacité pour la sous-population i

Les règles de décision dans le cas où on continue avec la sous-population 2 deviennent (Tableau 11) :

Tableau 11. ***Règles d'arrêt à la fin de l'étape 2 lorsque seule la sous-population 2 continue après l'étape 1***

Ω_2F2 = −1	Ω_2F2 = 1
$I_1 I_2$	$I_1 E_2$

I_i=Inefficacité pour la sous-population i ; E_i=Efficacité pour la sous-population i

Le nombre cumulé de patients inclus et le nombre cumulé de patients considérés pour la prise de décision finale sont décrits dans le Tableau 12.

Tableau 12. ***Nombre cumulé de réponses, nombre cumulé de patients inclus dans l'essai et nombre cumulé de patients utilisé pour la règle de décision finale, en fonction de la décision prise à l'étape 1***

Décision prise à l'étape 1	**Nombre cumulé de patients inclus dans l'essai**	**Nombre cumulé de réponses/ nombre cumulé de patients considérés pour la prise de décision finale**
Arrêt à l'étape 1	$N_{\cdot 1}$	$R_{\cdot 1} / N_{\cdot 1}$
Poursuite en étape 2 avec		
Sous-population 1 (Ψ_1= 1)	$N_{\cdot 1}+n_2$F1	$(r_{11}+r_2$F1$)/(n_{11}+n_2$F1$)$
Population entière (Ψ_1= 0)	$N_{\cdot 2}$	$R_{\cdot 2}/ N_{\cdot 2}$
Sous-population 2 (Ψ_1= 2)	$N_{\cdot 1}+n_2$F2	$(r_{21}+r_2$F1$)/(n_{21}+n_2$F2$)$

CALCULS

a) Effectifs et bornes d'arrêt

La méthode consiste à calculer 6 bornes d'arrêt (a_1, b_1, a_2, b_2, b_2F1, b_2F2) qui participeront à la définition des règles d'arrêt et 4 effectifs $N_{.1}$ et $N_{.2}$, n_2F1 et n_2F2.

Pour cela, on calcule :

π_0 : la probabilité marginale de réponse sous l'hypothèse nulle H_{00} : $\pi_1=\pi_{01}$ et $\pi_2=\pi_{02}$.
$\pi_0 = P(X = 1|i = 1) \times P(i = 1) + P(X = 1|i = 2) \times P(i = 2)$.

En s'appuyant sur la formule (1), on en déduit :

$$\pi_0 = \frac{\pi_{01} + \omega \times \pi_{02}}{1 + \omega}$$

De la même façon, on considère l'hypothèse alternative H_{11} : $\pi_1=\pi_{11}$ et $\pi_2=\pi_{12}$ pour calculer la probabilité marginale de réponse sous cette hypothèse,

$$\pi_1 = \frac{\pi_{11} + \omega \times \pi_{12}}{1 + \omega}.$$

Ce qui nous conduit à définir l'écart pondéré :

$$\Delta = \frac{\Delta_1 + \omega \times \Delta_2}{1 + \omega}.$$

On utilise la même méthode de calcul que le Plan de Fleming Classique pour calculer l'effectif nécessaire et on déduit les effectifs n_{11}, n_{21}, n_{12} et n_{22} de chaque sous-population i à chaque étape s de la formule (1). A noter que pour avoir un ratio entre les deux sous-populations strictement égal à ω, on impose ce même ratio lors des itérations.

Comme pour le Fleming classique les bornes a_s et b_s retenues sont celles qui permettent d'atteindre une puissance aussi proche (plus grande ou égale) que celle souhaitée, pour un risque α donné.

Dans l'hypothèse où seule la sous-population i participerait à l'étape 2, on utilise un plan de Fleming classique qui permet de calculer

- a_1Fi, b_1Fi, qui ne seront pas utilisés dans la méthode A
- n_2Fi et b_2Fi qui seront utilisés dans la méthode A

avec les contraintes suivantes :

- n_1Fi= n_{i1}
- αFi = α et βFi = β.

Un récapitulatif de la valeur des effectifs n_{is} et de b_2 à prévoir pour chaque sous-population en fonction de la décision qui sera observée à la fin de la première étape est présenté dans le Tableau 13.

Tableau 13. ***Valeur des effectifs n_{is}(i=1,2 et s=1,2) et de la borne d'arrêt à l'étape 2 de la méthode A pour chaque sous-population i en fonction de la décision prise à la fin de l'étape 1***

	***i*=1**	***i*=2**	b_2
Poursuite des 2 sous-populations	n_{12}	n_{22}	b_2
Arrêt des 2 sous-populations	0	0	-
Poursuite avec la sous-population 1	n_2F1	0	b_2F1
Poursuite avec la sous-population 2	0	n_2F2	b_2F2

b) Identification d'une hétérogénéité de réponses entre les sous-populations

Pour identifier une hétérogénéité de réponses entre les deux sous-populations, nous calculons un intervalle de fluctuation, IP_{is} = [inf_{is} ; sup_{is},], (i, s =1, 2), (Figure 1) au risque 1- γ , autour de chacun des π_{0i} des deux sous-populations. Nous pouvons définir ces intervalles soit à partir des pourcentages, soit à partir des effectifs. C'est cette seconde méthode que

nous utilisons ici et les bornes s'apparentent alors à des effectifs. Ces bornes sont des fonctions de γ :

$$sup_{is} = \min\left(c \middle| P(Y \geq c) < \frac{\gamma}{2}\right)$$

$$inf_{is} = \max\left(c \middle| P(Y \leq c) < \frac{\gamma}{2}\right)$$

avec Y suit une loi binomiale pour chaque sous-population i

- de paramètre (n_{i1}, π_{0i}) à l'étape 1
- de paramètre (n_{i1}+n_{i2}, π_{0i}) à l'étape 2.

Ces intervalles sont, par définition, symétriques en termes de probabilité autour de $n_{i1} \times \pi_{0i}$ ou de $(n_{i1} + n_{i2}) \times \pi_{0i}$.

La taille des échantillons n_{i1} étant en général petite, l'intervalle de probabilité ne peut pas être strictement égal à (1-γ) et seule une symétrie approximative peut être obtenue du fait du calcul binomial.

L'intervalle retenu est celui qui satisfait au mieux la condition de symétrie et de taille.

Nous proposons la définition de l'hétérogénéité suivante à la première étape : Les réponses entre les deux sous-populations sont considérées comme hétérogènes si (Figure 1) :

$$\frac{r_{11}}{n_{11}} \geq sup_{11} \text{ et } \frac{r_{21}}{n_{21}} \leq inf_{21} \ (\Psi_1 = 1) \text{ ou } \frac{r_{11}}{n_{11}} \leq inf_{11} \text{ et } \frac{r_{21}}{n_{21}} \geq sup_{21} \ (\Psi_1 = 2).$$

Dans toutes les autres situations, l'effet du traitement est considéré comme similaire entre les deux sous-populations et aucune sous-population n'est sélectionnée en particulier comme étant la seule sensible au traitement.

La probabilité de conclure à l'hétérogénéité des réponses observées à l'étape 1 est donnée par :

$$P(\Psi_1 = 1) + P(\Psi_1 = 2)$$
$$= \left(P\left(\frac{r_{11}}{n_{11}} \geq sup_{11}\right) \times P\left(\frac{r_{21}}{n_{21}} \leq inf_{21}\right) + P\left(\frac{r_{11}}{n_{11}} \leq inf_{11}\right) \times P\left(\frac{r_{21}}{n_{21}} \geq sup_{21}\right) \right) \leq \frac{\gamma^2}{2}.$$

La probabilité de conclure à l'hétérogénéité des réponses observées à l'étape 2 est donnée par :

$$P(\Psi_2 = 1) + P(\Psi_2 = 2)$$
$$= \left(P\left(\frac{R_{1\cdot}}{N_{1\cdot}} \geq sup_{12}\right) \times P\left(\frac{R_{2\cdot}}{N_{2\cdot}} \leq inf_{22}\right) + P\left(\frac{R_{1\cdot}}{N_{1\cdot}} \leq inf_{12}\right) \times P\left(\frac{R_{2\cdot}}{N_{2\cdot}} \geq sup_{22}\right) \right) \leq \frac{\gamma^2}{2}.$$

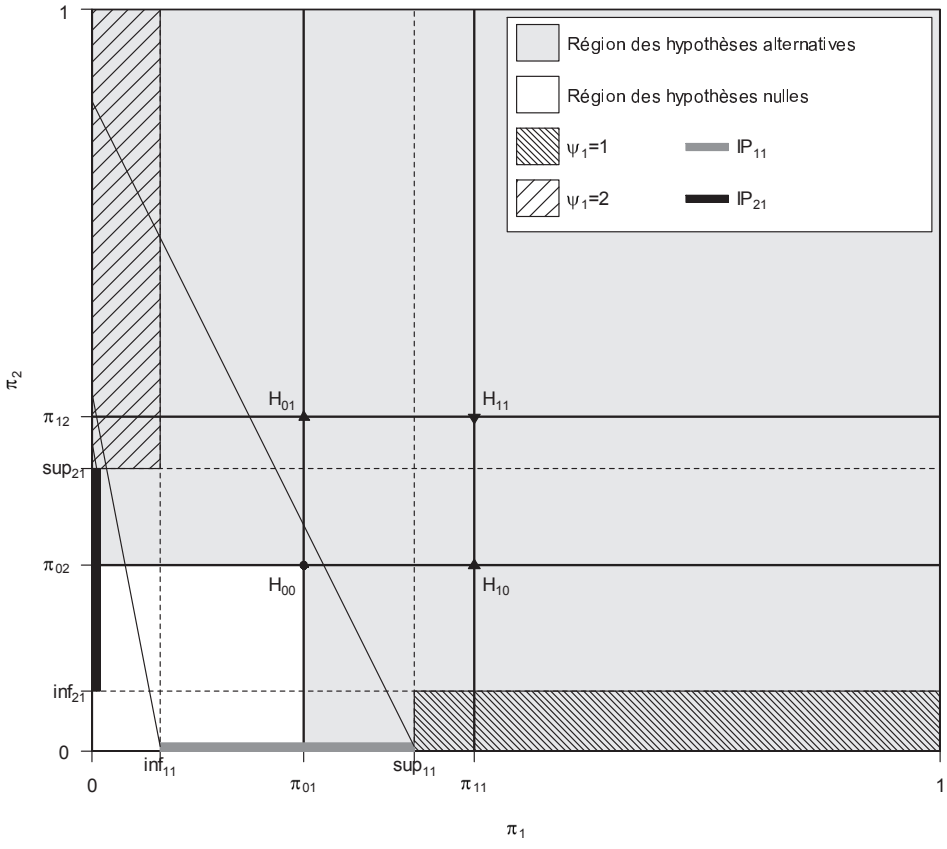

Figure 1. Représentation graphique des d'hypothèses nulles et alternatives, des hypothèses particulières H_{00}, H_{01}, H_{10} et H_{11}, de la détection de l'hétérogénéité de réponses à la première étape (Ψ_1=1 ou Ψ_1=2) et des intervalles de probabilité IP_{i1}

c) Détermination des erreurs de type I et II

Le plan de Fleming classique utilisé pour construire la méthode A satisfait les contraintes sur les risques d'erreurs de type I et II choisis dans le protocole et notés α et β.

Dans la méthode A, les risques calculés sont donnés par :

$\alpha_A = \mathrm{P}\{\Phi = 1 \mid \pi_1 = \pi_{01} \text{ et } \pi_2 = \pi_{02}\}$

$\beta_A = \mathrm{P}\{\Phi = 0 \mid \pi_1 = \pi_{11} \text{ et } \pi_2 = \pi_{12}\}$

Notons que du fait de l'ajout de la règle d'hétérogénéité, ils peuvent ne plus être inférieurs à α et β.

Le Tableau 14 présente les différentes combinaisons de décisions possibles suite à l'étape 1 puis 2 :

- L'essai peut être interrompu dès l'étape 1 et la conclusion finale est :
  - ✓ Soit efficacité démontrée pour les deux sous-populations, soit efficacité pour l'une et inefficacité pour l'autre sous-population car une hétérogénéité dans les réponses a été détectée (cas A)
  - ✓ Soit inefficacité sur l'ensemble de la population (cas C)

- L'essai peut se poursuivre en 2ème étape avec au moins une sous-population et la conclusion finale est :
  - ✓ Soit efficacité démontrée pour les deux sous-populations, soit efficacité pour l'une et inefficacité pour l'autre sous-population car une hétérogénéité dans les réponses a été détectée, parfois dès l'étape 1 (cas B_1, B_2, quand seule une sous-population poursuit en étape 2 et B_{12} quand les deux sous-populations poursuivent en étape 2)
  - ✓ Soit inefficacité sur l'ensemble de la population (cas D_1, D_2, quand seule une sous-population poursuit en étape 2 et D_{12} quand les deux sous-populations poursuivent en étape 2)

Tableau 14. Combinaisons de décisions possibles suite aux étapes 1 et 2

			Décisions à la fin de l'étape 1								
		Ψ_1	I_1I_2	C_1I_2	I_1C_2	C_1C_2	C_1I_2	I_1C_2	E_1E_2	E_1I_2	I_1E_2
Décisions à la fin de l'étape 2	I_1I_2	0				D_{12}					
	I_1I_2	1		D_1			D_1				
	I_1I_2	2			D_2			D_2			
	E_1E_2	0				B_{12}					
	E_1I_2	0				B_{12}					
	I_1E_2	0				B_{12}					
	E_1I_2	1		B_1			B_1				
	I_1E_2	2			B_2			B_2			
Arrêt étape 1		0	C						A		
		1								A	
		2									A

I=Inefficacité (de la sous-population 1, I_1 ou 2, I_2) ; E=Efficacité (de la sous-population 1, E_1 ou 2, E_2) ; C=Poursuite à l'étape 2 (avec la sous-population 1, C_1 ou 2, C_2).

Pour chaque situation, il est possible de calculer une probabilité de survenue sous H_{00} ou sous H_{11} et donc d'en déduire les probabilités α_A et β_A .

Ainsi, on a :

- $\alpha_A = P\{\Phi = 1 \mid \pi_1 = \pi_{01} \text{ et } \pi_2 = \pi_{02}\} = A(H_{00}) + B_{12}(H_{00}) + B_1(H_{00}) + B_2(H_{00})$
- $\beta_A = P\{\Phi = 0 \mid \pi_1 = \pi_{11} \text{ et } \pi_2 = \pi_{12}\} = C(H_{11}) + D_{12}(H_{11}) + D_1(H_{11}) + D_2(H_{11})$

EVALUATION DES CARACTERISTIQUES OPERATOIRES

a) Calculs théoriques

Tous les résultats sont obtenus en réalisant des calculs exacts et en utilisant la loi binomiale.

Nous avons étudié plusieurs scenarii correspondant à différents "vrais" taux de succès pour les deux sous-populations variant entre 0 et 1 par pas de 0.01, afin d'évaluer les caractéristiques opératoires de la méthode A.

Ainsi sont déterminés :

- les effectifs maximaux à inclure :

$$N_{max} = max\left(\sum_{i=1}^{2}\sum_{s=1}^{2} n_{is}\,,\ \sum_{i=1}^{2} n_{i1} + n_2F1,\ \sum_{i=1}^{2} n_{i1} + n_2F2\right)$$

- les effectifs attendus sous les hypothèses nulle et alternatives :

$$E(n|H_{00}),\ E(n|H_{01}),\ E(n|H_{10}),\ E(n|H_{11})$$

avec n l'effectif total de patients à inclure dans l'essai ($n = N_2$) et

$E(n|H) = N_{\cdot 1}$

$+ \left(P(D_1|H) + P(B_1|H)\right) \times n_2 F1$

$+ (P(D_2|H) + P(B_2|H)) \times n_2 F2$

$+ (P(D_{12}|H) + P(B_{12}|H)) \times (n_{12} + n_{22})$

pour $H = H_{00}, H_{01}, H_{10}$ ou H_{11}.

Le taux de bonnes conclusions finales, la probabilité de conclure à l'inefficacité ou à l'efficacité sur l'ensemble de la population sous H_{01} ou H_{10}, la probabilité de détecter une hétérogénéité de réponses entre les deux sous-populations à la première étape sous H_{01}, H_{10} ou H_{11} et la probabilité de continuer en phase III sous H_{00} (α_A) ou H_{11} (β_A) sont également calculés.

Les résultats obtenus avec notre méthode A sont comparés à ceux obtenus avec un schéma de Fleming classique (méthode H).

Par simplification, seuls quelques exemples proches des situations rencontrées en pratique clinique seront présentés, à savoir :
$\pi_{01} = \pi_{02}$, $\Delta_1 = \Delta_2$, w=1, α =0.05, β =0.1 et γ=0.6, et pour π_{0i}=0.25, γ =0.3 et 0.8.

On note que pour γ=0.3, 0.6 et 0.8, on obtient une probabilité de déclarer une hétérogénéité sous H_{00} d'approximativement $\frac{\gamma^2}{2}$ soit 4.5%, 18% et 32%.

b) Illustration à partir d'une problématique réelle

Notre méthode est illustrée par un exemple basé l'essai REMAGUS 02 (Pierga, 2009).

L'objectif de cette étude était de déterminer l'efficacité de deux nouveaux traitements, Trastuzumab et Celecoxib, en combinaison avec la chimiothérapie néoadjuvante standard, pour des patientes atteintes de cancer du sein et surexprimant HER_2 ou pas, le statut HER_{2+} ayant un pronostic plus défavorable. La planification de l'étude était la suivante :

- Les patientes HER_{2+} étaient randomisées entre Traitement Standard *versus* Traitement Standard+Trastuzumab.
- Les patientes HER_{2-} étaient randomisées entre Traitement Standard *versus* Traitement Standard+Celecoxib.

Cet essai a été planifié selon deux plans de Fleming à deux étapes, conduits en parallèle : l'un mené chez les femmes HER_{2+} et l'autre chez les femmes HER_{2-}.

Pour les besoins de ce travail, Trastuzumab et Celecoxib seront appelés "Traitement expérimental" et on ne considèrera que les femmes des deux bras recevant ces nouveaux traitements (et donc que l'objectif est de déterminer l'efficacité de la combinaison « Traitement Standard + Traitement expérimental » chez les femmes HER_{2+} et HER_{2-} (qui sont deux sous-populations des femmes atteintes de cancer du sein).

Donc, pour cet exemple les femmes randomisées dans les bras Traitement Standard seul ne sont pas prises en compte.

Les paramètres des schémas retenus dans les plans de Fleming initiaux sont :

- Sous-population des femmes HER_{2+} :
  $H_0 : \pi_1 \leq 0.15$ & $H_1 : \pi_1 > 0.15$ (avec un taux d'efficacité optimale de 0.30)
- Sous-population des femmes HER_{2-} :
  $H_0 : \pi_2 \leq 0.15$ & $H_1 : \pi_2 > 0.15$ (avec un taux d'efficacité optimale de 0.25)

En effet, les taux de réponse publiés en population non sélectionnée sur le statut HER_2 avec le Traitement Standard sont de 15%. Dans le contexte de essais, l'efficacité du Traitement expérimental est considérée comme cliniquement intéressante si ce dernier induit un taux de réponses d'au moins 30% dans la sous-population des femmes HER_{2+} et 25% dans la sous-population des femmes HER_{2-}.

Dans le protocole et pour des considérations de tailles d'échantillon, les risques α et β étaient fixés à 0.07 et 0.10 pour la sous-population HER_{2+} et 0.09 et 0.10 pour la sous-population HER_{2-}. Or avec la méthode que nous proposons (A), on ne peut proposer qu'un risque α et β pour l'ensemble des deux sous-populations ; dans un souci d'homogénéisation et de simplification les risques α et β ont donc été fixés à 0.05 et 0.10 respectivement, conduisant à la nécessité de simuler les réponses de patients supplémentaires en utilisant les taux de réponses observés lors de la conduite de l'essai pour le calcul binomial.

Comme environ 20 à 30 % des femmes sont HER_{2+}, nous avons considéré que $w = 3$ dans cet exemple soit 25% de femmes HER_{2+}.

Nous avons également choisi une valeur de γ égale 0.6.

RESULTATS

RESULTATS THEORIQUES

a) Adéquation des méthodes A et H en termes d'effectifs (Tableau 15)

Les effectifs attendus des deux méthodes sous les hypothèses nulle H_{00} ($E(n|H_{00})$) ou alternative H_{11} ($E(n|H_{11})$) sont très similaires.
Par exemple, quand π_{0i}=0.25, $E(n|H_{00})$ est égal à 38.46 avec la méthode H et 38.74, 39.88 ou 40.73 avec la méthode A pour des valeurs de γ=0.3, 0.6 ou 0.8, respectivement.

Cela signifie que lorsque la nouvelle thérapeutique est inefficace ou à l'inverse efficace pour l'ensemble de la population (absence d'hétérogénéité), alors la méthode A est similaire à la méthode H en terme d'effectifs attendus.

On constate en revanche que lorsqu'une seule sous-population est effectivement sensible à la nouvelle thérapeutique (présence d'une hétérogénéité), les effectifs attendus $E(n|H_{01})$ ou $E(n|H_{10})$ augmentent avec la méthode A, de près de 10%.
Il faut cependant souligner que ces patients supplémentaires, inclus dans l'essai conduit selon la méthode A sont des patients issus de la sous-population sensible à la nouvelle thérapeutique. L'augmentation de leur nombre est donc éthiquement acceptable.

Ceci a dans le même temps un impact sur l'effectif maximal pouvant être inclus qui est alors supérieur avec la méthode A par rapport à la méthode H.

Tableau 15. ***Effectifs maximaux et attendus avec les méthodes A et H sous les hypothèses nulle, alternative, ou combinée pour des taux de réponse identiques sous l'hypothèse nulle, Δi=0.2, w=1, α=0.05 et β=0.1***

π_{0i}	γ	Effectif maximal		Effectif attendu					
		A	H	A			H		
				H_{00} [§]	H_{01} ou H_{10}	H_{11}	H_{00}	H_{01} ou H_{10}	H_{11}
0.05	**0.6**	32	28	21.11	24.77	21.05	21.11	24.77	21.05
0.10	**0.6**	42	36	27.78	32.09	27.36	27.78	32.09	27.36
0.15	**0.6**	48	40	32.35	37.56	31.93	31.78	36.39	31.78
0.20	**0.6**	57	48	38.11	44.23	35.70	36.66	41.95	35.42
0.25	**0.3**	62	52	**38.74**	47.29	41.86	**38.46**	46.43	41.80
0.25	**0.6**	62	52	**39.88**	48.23	41.93	**38.46**	46.43	41.80
0.25	**0.8**	62	52	**40.73**	49.85	42.27	**38.46**	46.43	41.80
0.30	**0.6**	67	56	43.09	51.84	43.78	41.01	49.27	43.59
0.35	**0.6**	69	56	41.34	52.49	45.98	38.62	48.33	45.59
0.40	**0.6**	70	56	42.25	52.59	44.63	40.34	48.86	44.24
0.45	**0.6**	68	56	41.26	52.02	46.67	38.08	48.05	46.41
0.50	**0.6**	67	56	42.51	52.75	45.46	39.73	48.96	45.20
0.55	**0.6**	67	56	40.26	52.99	48.12	37.46	48.02	47.79
0.60	**0.6**	60	48	34.25	44.39	40.97	31.79	40.57	40.80
0.65	**0.6**	53	44	30.43	40.42	40.29	28.63	37.18	40.18
0.70	**0.6**	47	40	25.94	35.41	36.77	24.73	32.42	36.70
0.75	**0.6**	38	32	19.66	26.59	31.32	19.15	24.93	31.31

[§]*Sous H_{00}, $\pi_1 = \pi_{01}$ & $\pi_2 = \pi_{02}$; sous H_{01}, $\pi_1 = \pi_{01}$ & $\pi_1 = \pi_{12}$; sous H_{10}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{02}$; sous H_{11}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{12}$*

b) Amélioration de la pertinence des décisions avec la méthode A

Augmentation du taux de bonnes conclusions finales (Tableau 16)

L'un des avantages clé de la méthode A est le gain dans le nombre de bonnes conclusions conduisant à la poursuite de l'évaluation de la nouvelle thérapeutique sur une sous-population sensible quand l'évaluation aurait été arrêtée pour l'ensemble de la population avec la méthode H.

En effet, par définition, quand une seule sous-population est sensible à la nouvelle thérapeutique (H_{01} ou H_{10}), la méthode H est incapable de l'identifier puisqu'elle n'a pas été élaborée dans ce but, contrairement à la méthode A.

Par exemple, avec la méthode A, une bonne conclusion (Inefficacité pour la sous-population 1 et Efficacité pour la sous-population 2 sous H_{01} ou à

l'inverse Efficacité pour la sous-population 1 et Inefficacité pour la sous-population 2 sous H_{10}) est donnée dans 10.7% à 21.7% des cas (pour des π_{0i} compris entre 0.15 et 0.70, et jusqu'à 26.3% selon les hypothèses posées sur γ).

Cette capacité peut sembler modeste, mais elle représente une amélioration non négligeable par rapport à la méthode classique qui n'aurait pas pu faire de distinction et donc un potentiel important de thérapeutiques pour lesquelles l'évaluation serait poursuivie alors qu'elle est abandonnée si l'on suit les conclusions de la méthode de Fleming classique.

Tableau 16. ***Taux de bonnes conclusions finales des méthodes A et H sous les hypothèses nulle H_{00}, alternative H_{11}, ou combinée H_{01} et H_{10} pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, ω=1, α=0.05 et β=0.1***

π_{0i}	γ	H_{00} [§]		H_{01}/H_{10}		H_{11}	
		A	**H**	**A**	**H**	**A**	**H**
0.05	**0.6**	0.953	0.953	0.000	0	0.936	0.936
0.10	**0.6**	0.974	0.974	0.018	0	0.886	0.887
0.15	**0.6**	0.958	0.968	0.145	0	0.866	0.878
0.20	**0.6**	0.937	0.952	**0.217**	0	0.896	0.917
0.25	**0.3**	0.953	0.957	0.072	0	0.906	0.910
0.25	**0.6**	0.950	0. 957	0.104	0	0.906	0.910
0.25	**0.8**	0.941	0. 957	**0.263**	0	0.881	0.910
0.30	**0.6**	0.939	0.948	0.135	0	0.919	0.924
0.35	**0.6**	0.939	0.951	0.163	0	0.901	0.910
0.40	**0.6**	0.938	0.950	0.195	0	0.895	0.911
0.45	**0.6**	0.946	0.956	0.135	0	0.898	0.902
0.50	**0.6**	0.946	0.958	0.167	0	0.902	0.908
0.55	**0.6**	0.956	0.966	0.186	0	0.898	0.905
0.60	**0.6**	0.949	0.958	0.122	0	0.902	0.903
0.65	**0.6**	0.966	0.973	0.110	0	0.876	0.877
0.70	**0.6**	0.970	0.977	**0.107**	0	0.887	0.887
0.75	**0.6**	0.973	0.977	0.074	0	0.910	0.910

[§] *Sous H_{00}, $\pi_1=\pi_{01}$ & $\pi_2=\pi_{02}$; sous H_{01}, $\pi_1=\pi_{01}$ & $\pi_1=\pi_{12}$; sous H_{10}, $\pi_1=\pi_{11}$ & $\pi_1=\pi_{02}$; sous H_{11}, $\pi_1=\pi_{11}$ & $\pi_1=\pi_{12}$*

Diminution du taux d'arrêt complet du développement pour inefficacité pour des sous-populations sensibles (Tableau 17)

La probabilité de conclure à l'inefficacité pour l'ensemble de la population alors que le vrai taux de réponse dans les sous-populations n'est égal à π_{0i} que pour l'une d'entre elles est inférieure avec la méthode A par rapport à la méthode H.

Par exemple, sous H_{01} ou H_{10}, quand 51.6% des conclusions finales de l'essai sont « Inefficacité de la nouvelle thérapeutique pour l'ensemble de la population » avec la méthode H, ce taux n'est que de 38.1% avec la méthode A (π_{0i}=0.2).

Cela signifie là encore que moins de développements de nouvelles thérapeutiques sont arrêtés en phase II quand qu'une partie de la population peut en bénéficier.

Tableau 17. ***Probabilité de conclure à l'Inefficacité pour la population entière avec les méthodes A et H sous les combinée H_{01} et H_{10} pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, ω=1, α=0.05 et β=0.1***

π_{0i}	γ	H_{01}/H_{10}	
		A	**H**
0.05	**0.6**	0.381	0.381
0.10	**0.6**	0.562	0.562
0.15	**0.6**	0.482	0.582
0.20	**0.6**	**0.381**	**0.516**
0.25	**0.3**	0.495	0.545
0.25	**0.6**	0.472	0.545
0.25	**0.8**	0.381	0.545
0.30	**0.6**	0.427	0.519
0.35	**0.6**	0.428	0.549
0.40	**0.6**	0.423	0.558
0.45	**0.6**	0.476	0.588
0.50	**0.6**	0.465	0.598
0.55	**0.6**	0.472	0.629
0.60	**0.6**	0.510	0.622
0.65	**0.6**	0.598	0.706
0.70	**0.6**	0.627	0.734
0.75	**0.6**	0.681	0.755

[§]*Sous H_{01}, $\pi_1 = \pi_{01}$ & $\pi_1 = \pi_{12}$; sous H_{10}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{02}$*

Diminution du taux de poursuite du développement pour efficacité pour des sous-populations non sensibles (Tableau 18)

La probabilité de conclure à l'efficacité pour l'ensemble de la population alors que le vrai taux de réponse dans les sous-populations est égal à π_{0i} ou π_{10} est inférieure avec la méthode A par rapport à la méthode H.

Par exemple, pour π_{0i}=0.2, quand 48.4% des conclusions d'essai sont « Efficacité pour l'ensemble de la population » avec la méthode H, ce taux n'est que de 40.1% avec la méthode A.

Tableau 18. ***Probabilité de conclure à l'Efficacité pour la population entière avec les méthodes A et H sous les combinée H_{01} et H_{10} pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, ω=1, α=0.05 et β=0.1***

π_{0i}	γ	H_{01}/H_{10}	
		A	**H**
0.05	**0.6**	0.619	0.619
0.10	**0.6**	0.420	0.438
0.15	**0.6**	0.373	0.418
0.20	**0.6**	**0.401**	**0.484**
0.25	**0.3**	0.433	0.455
0.25	**0.6**	0.424	0.455
0.25	**0.8**	0.355	0.455
0.30	**0.6**	0.438	0.481
0.35	**0.6**	0.408	0.451
0.40	**0.6**	0.382	0.442
0.45	**0.6**	0.389	0.412
0.50	**0.6**	0.368	0.402
0.55	**0.6**	0.341	0.371
0.60	**0.6**	0.368	0.378
0.65	**0.6**	0.292	0.294
0.70	**0.6**	0.266	0.266
0.75	**0.6**	0.245	0.245

§*Sous H_{01}, $\pi_1 = \pi_{01}$ & $\pi_1 = \pi_{12}$; sous H_{10}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{02}$*

Cela se traduit concrètement par un bénéfice éthique puisque moins de patients non sensibles à la nouvelle thérapeutique y seront exposés : ils ne participeront pas à la phase III et ne subiront donc pas à tort les potentiels effets toxiques de la nouvelle thérapeutique d'une part, et d'autre part, ils ne risqueront pas d'entraîner une dilution de l'effet traitement en phase III, aboutissant à un échec apparent de la nouvelle thérapeutique alors qu'en réalité elle pourrait bénéficier à une partie de la population.

Détection non négligeable de l'hétérogénéité dès la première étape (Tableau 19)

La méthode A est d'autant plus intéressante d'un point de vue éthique qu'elle peut détecter la population non-sensible dès la fin de la première étape (jusqu'à 21.7% selon les hypothèses), réduisant encore ce nombre de patients exposés à tort à une nouvelle thérapeutique potentiellement toxique et dont la poursuite de participation à l'essai retarde la prise en charge thérapeutique puisqu'ils ne peuvent pas bénéficier d'un autre traitement tant qu'ils participent à cette recherche.

Si la probabilité de ne détecter qu'une seule sous-population sensible alors que l'ensemble de la population est sensible n'est pas formellement contrôlée dans la méthode A, on peut noter que cette probabilité (ou proportion de thérapeutiques jugées "Efficaces pour une sous-population et Inefficace pour l'autre sous-population" sous H_{11}) est faible et varie à la première étape de 0 à 0.03 quand par exemple π_{0i}=0.25 et γ=0.6.

Enfin la probabilité de détecter une hétérogénéité sous H_{00} varie entre 0 et 11.3% (à comparer à $\frac{\gamma^2}{2}$ = 18%).

Tableau 19. ***Détection de l'hétérogénéité à l'étape 1 avec la méthode A sous les hypothèses nulle H_{00}, alternative H_{11}, ou combinée H_{01} et H_{10} pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, ω=1, α=0.05 et β=0.1***

π_{0i}	γ	H_{00} [§]	H_{01}/H_{10}	H_{11}
		A	**A**	**A**
0.05	**0.6**	0	0	0
0.10	**0.6**	0	0	0
0.15	**0.6**	0.071	0.148	0.020
0.20	**0.6**	**0.113**	**0.217**	**0.030**
0.25	**0.3**	0.020	0.073	0.006
0.25	**0.6**	0.052	0.099	0.008
0.25	**0.8**	0.137	0.262	0.042
0.30	**0.6**	0.071	0.128	0.010
0.35	**0.6**	0.081	0.166	0.017
0.40	**0.6**	0.084	0.196	0.024
0.45	**0.6**	0.086	0.138	0.010
0.50	**0.6**	0.090	0.167	0.013
0.55	**0.6**	0.086	0.194	0.015
0.60	**0.6**	0.071	0.127	0.006
0.65	**0.6**	0.060	0.116	0.004
0.70	**0.6**	0.045	0.111	0.002
0.75	**0.6**	0.023	0.076	0

[§]*Sous H_{00}, $\pi_1=\pi_{01}$ & $\pi_2=\pi_{02}$; sous H_{01}, $\pi_1=\pi_{01}$ & $\pi_1=\pi_{12}$; sous H_{10}, $\pi_1=\pi_{11}$ & $\pi_1=\pi_{02}$; sous H_{11}, $\pi_1=\pi_{11}$ & $\pi_1=\pi_{12}$*

Contrôle satisfaisant des risques d'erreur (Tableau 20)

Le risque d'erreur de type I est bien contrôlé. Néanmoins, il augmente légèrement quand γ augmente.

Si l'on considère l'exemple suivant : π_{0i}=0.25 et γ=0.3, 0.6 ou 0.8, les probabilités de continuer à tort en phase III alors qu'aucune sous-population n'est en réalité sensible à la nouvelle thérapeutique (H_{00}) sont de 0.047, 0.050 et 0.059 respectivement alors que le seuil α fixé par le protocole était de 0.05.

La puissance associée à la méthode A est satisfaisante pour toutes les valeurs de π_{0i}.

Ainsi, sur l'ensemble des valeurs, on constate que la probabilité de poursuivre l'évaluation en phase III quand l'ensemble de la population peut bénéficier de la thérapeutique varie entre 0.880 et 0.926.

Tableau 20. ***Probabilité de passer en phase III avec les méthodes A et H sous les hypothèses nulle H_{00}, ou alternative H_{11}, pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, ω=1, α=0.05 et β=0.1***

π_{0i}	γ	H_{00} [§]		H_{11}	
		A	**H**	**A**	**H**
0.05	**0.6**	0.047	0.047	0.936	0.936
0.10	**0.6**	0.026	0.026	0.887	0.887
0.15	**0.6**	0.042	0.032	0.885	0.878
0.20	**0.6**	0.063	0.048	**0.926**	0.917
0.25	**0.3**	**0.047**	0.043	0.912	0.910
0.25	**0.6**	**0.050**	0.043	0.914	0.910
0.25	**0.8**	**0.059**	0.043	0.921	0.910
0.30	**0.6**	0.061	0.052	0.929	0.924
0.35	**0.6**	0.061	0.049	0.918	0.910
0.40	**0.6**	0.062	0.050	0.919	0.911
0.45	**0.6**	0.054	0.044	0.908	0.902
0.50	**0.6**	0.054	0.042	0.914	0.908
0.55	**0.6**	0.044	0.034	0.913	0.905
0.60	**0.6**	0.051	0.042	0.908	0.903
0.65	**0.6**	0.034	0.027	**0.880**	0.877
0.70	**0.6**	0.030	0.023	0.890	0.887
0.75	**0.6**	0.027	0.023	0.911	0.910

[§] *Sous H_{00}, $\pi_1 = \pi_{01}$ & $\pi_2 = \pi_{02}$; sous H_{01}, $\pi_1 = \pi_{01}$ & $\pi_1 = \pi_{12}$; sous H_{10}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{02}$; sous H_{11}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{12}$*

Quand les paramètres de la méthode A ne sont pas équilibrés (taux de réponse différents sous l'hypothèse nulle, Δ_1 différent de Δ_2, ω différent de 1, tous ces résultats sont retrouvés de façon similaire (données non présentées).

a) Adéquation des résultats théoriques avec l'application pratique en termes d'effectifs

Dans l'essai REMAGUS 02, deux Fleming indépendants ont été menés en parallèle. Parce que le risque α a été fixé à 0.05, 2 réponses supplémentaires ont été simulées dans la sous-population des femmes HER_{2+} et 28 dans la sous-population de femmes HER_{2-}.

Avec ce schéma d'étude, 64 femmes de la sous-population 1, *i. e.* avec récepteurs HER_{2+}, sont à prévoir, soit 32 à chaque étape ; contre 136 femmes issues de la sous-population 2, *i. e.* avec récepteurs HER_{2-}, soit 68 à chaque étape. Au total, 200 femmes sont susceptibles de participer à l'essai REMAGUS 02 conduit selon deux plans de Fleming menés en parallèle (Tableau 21).

Tableau 21. ***Paramètres des schémas d'étude tels qu'appliqués lors de l'essai REMAGUS 02 (Deux Fleming indépendants conduits en parallèle ou méthode D) et suivant les méthodes A et H.***

Méthode			**A**		**H**	**D**	
Population			**HER_{2+}**	**HER_{2-}**	**HER_{2+}/HER_{2-}**	**HER_{2+}**	**HER_{2-}**
π_{0i} - π_{1i}			0.15-0.30	0.15-0.25	0.15-0.26	0.15-0.30	0.15-0.25
α			0.05		0.05	0.05	0.05
β			0.10		0.10	0.10	0.10
w			3		3	-	-
γ			0.6		-	-	-
Effectifs par étape	**Etape 1**	n_{i1}	14	42	56	32	68
	Etape 2	n_{i2}	14	42	56	32	68
		n_2**Fi**	50	94	-	-	-
Effectif cumulé maximal			**150**		**112**	**200**	
Bornes	**Etape 1**	$]a_1 - b_1[$	8-16		8-16	4 - 10	9 - 18
	Etape 2	$\geq b_2$	24		24	15	28
		b_2**Fi**	15	28	-	-	

Comme attendu suite aux résultats théoriques décrits plus haut, on constate que si l'essai était conduit selon la méthode A, le recrutement nécessaire pour l'essai serait un peu plus important qu'avec une méthode utilisée dans un essai sans aucune distinction entre les deux sous-populations (essai suivant un seul plan de Fleming classique ou méthode H).

Ainsi, l'effectif maximal avec la méthode A serait de 150 patientes (14+42+94 dans l'hypothèse où seule la sous-population des femmes HER_{2-} continuerait en étape 2) contre 112 patientes avec la méthode H. Mais on rappelle que la méthode H ne permettrait pas de répondre à la question posée sur l'interaction entre le statut vis-à-vis des récepteurs HER_2 et l'effet du traitement.

Par contre, ces effectifs seraient évidemment beaucoup plus faibles qu'avec le schéma suivi dans la réalité, à savoir deux plans de Fleming classiques (D) conduits en parallèle.

Les effectifs attendus seraient très similaires pour les méthodes H et A, sauf quand une seule sous-population peut bénéficier du traitement où l'on retrouve une légère augmentation du nombre de patients (de l'ordre de 5%), résultat confirmant également les résultats théoriques.

La Figure 2 représente la différence entre les méthodes A et H (A-H) du nombre de patients inclus en fonction du vrai taux de réponse dans chacune des deux sous-populations.

On constate (Quadrants supérieur gauche et inférieur droit où les vrais taux de réponses sont supérieurs à π_{0i}) que les patients supplémentaires inclus à l'étape 2 quand la méthode A est suivie sont ceux qui sont bien sensibles à la nouvelle thérapeutique et qu'ils sont au nombre de 5 à 25 de plus.

On voit également dans cet exemple que, si le vrai taux de réponse de la sous-population 1 est très largement supérieur à l'hypothèse faite sous H_1

alors le nombre de sujets inclus est même inférieur avec la méthode A puisqu'il est possible avec cette méthode de s'arrêter dès l'étape 1 pour efficacité dans une seule sous-population.

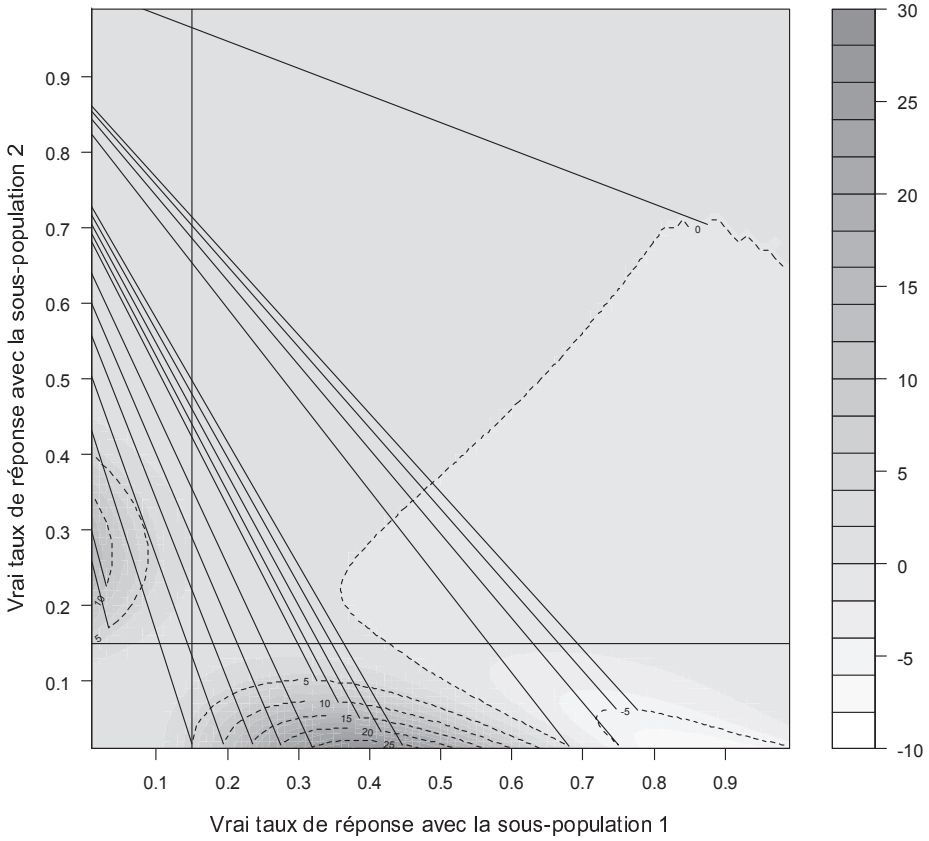

Figure 2. Différence du nombre de patients inclus avec la méthode A par rapport à la méthode H en fonction du vrai taux de réponse des deux sous-populations 1 et 2. Les paramètres de cet exemple sont les suivants : α=0.08, β=0.1, γ=0.6, ω=3, $\pi_{01}=\pi_{02}$=0.15, π_{11}=0.3, π_{12}=0.25.

b) Adéquation des résultats théoriques avec l'application pratique en termes de pertinence des décisions

Sous H_{01} ou H_{10}, les taux de vraies conclusions avec la méthode A sont sous H_{01} de 0.39, 0.41 et 0.69 et sous H_{10} de 0.54, 0.64 et 0.69 pour des valeurs de γ de 0.3, 0.6 et 0.8.

La Figure 3 illustre le bénéfice en termes de conclusions faussement négatives (conclure à l'inefficacité pour l'ensemble de la population alors qu'une sous-population est sensible). On voit que lorsque que le vrai taux de réponses dans chaque sous-population se rapproche (ou *a fortiori* dépasse) l'hypothèse du taux de réponses sous l'hypothèse alternative, la probabilité de conclure à l'inefficacité pour l'ensemble de la population est bien moins importante avec la méthode A.

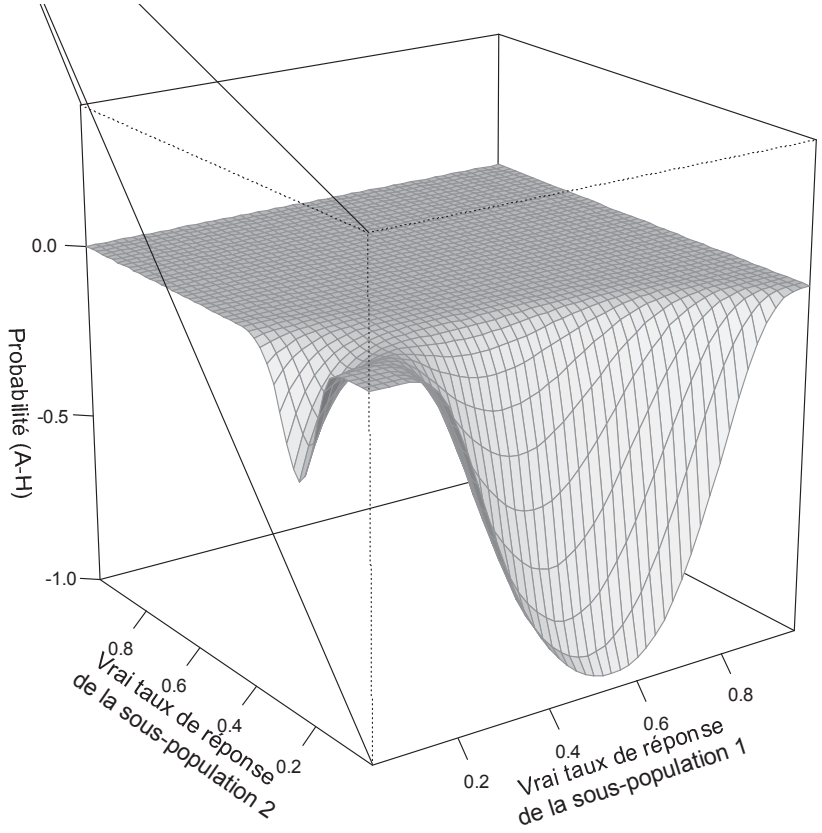

Figure 3. Différence entre la probabilité de conclure à l'Inefficacité pour l'ensemble de la population avec la méthode A comparativement à la méthode H (Probabilité (A-H)) en fonction des vrais taux de réponses de chaque sous-population i (i= 1, 2). Les paramètres de cet exemple sont les suivants : α=0.08, β=0.1, γ=0.6, ω=3, $\pi_{01}=\pi_{02}$=0.15, π_{11}=0.3, π_{12}=0.25.

c) Conclusion de la méthode A à partir des données observées dans l'essai REMAGUS 02 (Tableau 22)

Les conclusions qui auraient été obtenues si la nouvelle méthode A avait été utilisée sont comparés à ceux obtenus avec deux plans de Fleming conduits en parallèle comme cela a été le cas dans l'essai REMAGUS 02 (Méthode D) et à ceux qui auraient été obtenus si un plan de Fleming hétérogène avait été mené (méthode H).

Méthode D

Deux étapes pour chacun des plans ont été nécessaires pour conclure. Au total, 200 patientes sont donc évaluées : 64 femmes HER_{2+} et 136 femmes HER_{2-}. Avec les hypothèses fixées pour cet exemple l'efficacité des nouvelles thérapeutiques n'est démontrée que pour les femmes HER_{2+}.
C'était également la conclusion portée par les investigateurs de l'essai REMAGUS 02 avec les hypothèses qu'ils s'étaient fixées.

Méthode H

Avec un plan de Fleming classique, deux étapes sont également nécessaires et un total de 112 patientes (28 femmes HER_{2+} et 84 femmes HER_{2-}) participent à l'essai. La conclusion d'un essai mené selon cette approche aurait été que les nouvelles thérapeutiques étaient inefficaces. Ne faisant pas de distinction entre les sous-populations, la conclusion aurait été donnée pour l'ensemble de la population. A cause de la dilution de l'effet traitement, le développement des thérapeutiques aurait sans doute paru sans intérêt.

Méthode A

A l'étape 1, 5 réponses ont été observées parmi les 42 femmes HER_{2-} (soit 12% dans la sous-population 2) et 5 réponses parmi les 14 femmes HER_{2+} (soit 36% dans la sous-population 1). Avec 10 réponses à l'étape 1, $\Omega_1 = 0$. Les méthodes A, H ou D conduisent à la même conclusion à la fin de l'étape 1.

Comme aucune interaction n'est détectée à l'étape 1 (Ψ_1=0), toute la population poursuit en étape 2.
A l'étape 2, des taux de réponses cumulées de 11/84 (13%) chez les femmes (HER_{2+}) et de 9/28 (32%) chez les femmes HER_{2-} ont été observées. Pourtant, l'interaction n'a pas pu être observée car les intervalles IP_{12} et IP_{22} étaient de]2-6[et]10-15[respectivement. Le nombre de patientes inclus est de 112 patientes avec la méthode A, 112 et 200 avec les méthodes H et D respectivement.

La méthode A dans cet exemple n'aurait donc pas permis de mettre en avant l'efficacité des nouvelles thérapeutiques sur la sous-population des femmes HER_{2+} malgré un taux de réponse observé inférieur dans la sous-population HER_{2-} au taux de réponse sous l'hypothèse nulle (15%) et supérieur dans la sous-population HER_{2+} au taux de réponse sous l'hypothèse alternative (30%). Il est intéressant de souligner que si une réponse de moins avait été observée dans la sous-population 1, alors une interaction aurait été détectée.

Tableau 22. Hypothèses, règles de décision, résultats observés et conclusions finales en fonction du schéma d'étude suivi

		Méthode A		Méthode H	Méthode D	
Sous-Population*		1 : 25%	2 : 75%	1 : 25% 2 : 75%	1 : 100%	2 : 100%
Hypothèses		π_0 : 0.15 π_1 : 0.30	π_0 : 0.15 π_1 : 0.25	π_0 : 0.15 π_1 : 0.26	π_0 : 0.15 π_1 : 0.30	π_0 : 0.15 π_1 : 0.25
		α : 0.05 β : 0.10 γ : 0.6		α : 0.05 β : 0.10	α : 0.05 β : 0.10	α : 0.05 β : 0.10
Effectifs par étape	**Etape 1**	14	42	56 (14+42)	32	68
	Etape 2	14 ou 50	42 ou 94	56 (14+42)	32	68
Etape 1	**Réponses observées**	$\Omega_1 = 0$		$\Omega_1 = 0$	$\Omega_1 = 0$	$\Omega_1 = 0$
	Détection de l'hétérogénéité	$\Psi_1 = 0$				
	Décision	C_1C_2		C_1C_2	C_1	C_2
Etape 2	**Réponses observées**	$\Omega_2 = 0$		$\Omega_2 = 0$	$\Omega_2 = 1$	$\Omega_2 = 0$
Conclusion à la fin de l'essai		I_1I_2		I_1I_2	E_1	I_2
Nombre de participantes à l'essai		42+14+42+14=**112**		56+56=**112**	32+32+68+68=**200**	

**1= Femmes HER_2-, 2= Femmes HER_2+*
I=Inefficacité (de la sous-population 1, I_1 ou 2, I_2) ; E=Efficacité (de la sous-population 1, E_1 ou 2, E_2) ; C=Poursuite à l'étape 2 (avec la sous-population 1, E_1 ou 2, E_2).

SYNTHESE

La méthode que nous proposons dans cette partie permet d'étudier les taux de réponses observés au sein de deux sous-populations identifiées avant le début de l'étude. Ces deux sous-populations se distinguent sur une caractéristique connue avant le début de l'essai, qui peut être un marqueur biologique, une caractéristique clinique de la maladie ou de sa prise en charge.

L'objectif principal est avant tout d'éviter de conclure à l'inefficacité de la nouvelle thérapeutique pour l'ensemble des participants de l'étude alors qu'en réalité cette nouvelle thérapeutique peut avoir un intérêt pour l'une des deux sous-populations. Un autre objectif est d'éviter également que toute une population soit menée jusqu'en phase III alors qu'une partie d'entre elle n'est pas sensible au traitement. Cela pourrait arriver si en phase II la proportion de patients appartenant à cette sous-population est faible pour des raisons diverses. Le résultat observé ne serait alors pas représentatif de la sensibilité globale à la nouvelle thérapeutique mais uniquement le reflet de la grande sensibilité d'une seule sous-population, plus largement représentée au sein des participants.

Dans cette première approche, l'évaluation de l'hétérogénéité de réponses est basée sur la comparaison entre les taux de réponse observés et les bornes d'intervalles de probabilité construits autour des probabilités de réponse sous les hypothèses nulles.

Malgré les résultats intéressants présentés dans cette partie, il est apparu que la méthode peut manquer de puissance, en particulier quand les taux de réponses sous l'hypothèse nulle sont proches de 0 (comme dans l'essai REMAGUS 02) ou 1, conduisant alors à de larges intervalles de probabilité. Les résultats de la méthode sont également influencés par le choix du paramètre γ : le choix d'une valeur faible évite de détecter à tort une

hétérogénéité et donc de poursuivre l'évaluation avec l'une ou l'autre des deux sous-populations quand personne n'est sensible à la thérapeutique ; à l'inverse cela diminue dans le même temps la capacité à détecter précocement la sous-population sensible quand elle existe, rendant l'intérêt de la méthode quasi nul.

Le problème de cette première méthode est qu'elle ne tient pas compte de l'importance de la différence entre les taux de réponses observés entre les deux sous-populations. Ainsi comme dans l'exemple REMAGUS 02, les réponses peuvent ne pas être considérées comme hétérogènes car pour l'une des deux sous-populations le taux de réponse observé est juste en dedans des bornes de l'intervalle de probabilité alors qu'en pratique les deux taux de réponses observés sont vraiment différents d'un point de vue de clinique.

La capacité à détecter une hétérogénéité, bien qu'intéressante, pouvait donc sans doute être améliorée. Une réflexion a donc été entamée pour réduire cet écueil. C'est l'objet du chapitre suivant.

CHAPITRE 3. AMELIORATION DE LA METHODOLOGIE PROPOSEE ET APPLICATION AUX DONNEES REELLES DE L'ESSAI REMAGUS 02

Pour tenir compte de la différence entre les deux taux de réponses observées plutôt que de leur position par rapport aux intervalles de probabilité, nous proposons une nouvelle approche de l'identification de l'hétérogénéité permettant d'augmenter le taux de bonnes conclusions en cas d'hétérogénéité des réponses entre les deux sous-populations, sans nuire au risque de 1ère espèce.

Dans ce chapitre, nous présentons ce nouveau schéma de phase II (Méthode B) et l'appliquons à nouveau aux données de l'essai REMAGUS 02 pour mettre clairement en avant l'avantage de ce nouveau plan en pratique clinique. Des données théoriques issues comme au chapitre 2 de calculs binomiaux exacts sont également donnés.

Ce travail a fait l'objet d'une publication dans la revue Statistics in Medicine (Tournoux-Facon, 2011b) rapportée dans l'annexe 3.

PRESENTATION DE LA MODIFICATION PROPOSEE

NOTATIONS ET HYPOTHESES

Les mêmes notations et hypothèses que précédemment sont utilisées.

NOUVEAU PRINCIPE D'EVALUATION DE L'HETEROGENEITE

Cette amélioration de la méthode A a conduit à élaborer ce que nous appellerons par la suite la méthode B.

Avec la méthode B, on ne construit plus d'intervalles de probabilité autour des π_{0i}. L'idée est plutôt de s'intéresser à :

- l'importance de la différence pour chaque sous-population i entre les taux observés de réponses et les probabilités de réponses sous l'hypothèse nulle correspondante
- et au sens des différences (vont-elles dans le même sens ou pas ?)

Avec la méthode A, on déclarait à l'étape 1 une hétérogénéité en faveur de la sous-population 1 par exemple, si le pourcentage observé dans la sous-population 1 était suffisamment plus élevé que son π_{01} ($\pi_1 > sup_{11}$) et que le pourcentage observé dans la sous-population 2 était suffisamment plus bas que son π_{02} ($\pi_2 < inf_{21}$).

Avec la méthode B, si le pourcentage observé dans la sous-population 1 est plus grand strictement que π_{01} (et pas nécessairement supérieur à sup_{11}), l'hétérogénéité sera déclarée en faveur de la sous population 1 si et seulement si le pourcentage observé dans la sous-population 2 est strictement intérieur à son π_{02} (et pas nécessairement inférieur à inf_{21}) et que la somme des deux différences dépasse une certaine limite, limite que nous définissons dans le paragraphe suivant.

C'est le seul point méthodologiquement différent entre les méthodes A et B.

La détermination de l'hétérogénéité est comme précédemment notée Ψ_s.

A la fin de l'étape 1, la statistique Ψ_1 est toujours combinée à la statistique Ω_1 pour conduire aux mêmes conclusions que celles décrites dans le Tableau 8.

Les mêmes règles quant à l'adaptation des effectifs de l'étape 2 et de la borne d'arrêt b_2 sont retenues dans la situation où seule une sous-population est conduite en étape 2.

Les mêmes formules sont développées pour les erreurs de type I et II que dans le chapitre 2.

CALCUL

Pour identifier une éventuelle hétérogénéité entre les réponses des deux sous-populations, nous introduisons les statistiques suivantes :

$$d_s = \sum_{i=1}^{2} |d_{is}|$$

$$S_s = \sum_{i=1}^{2} sign(d_{is}),$$

avec :

- $d_{is} = \frac{R_{is}}{N_{is}} - \pi_{0i}$, différence pour chaque sous-population i entre les taux observés de réponses et les probabilités de réponses sous l'hypothèse nulle correspondante

- sign(x) égal à 1, 0 ou -1 selon que x est supérieur à 0, égal à 0 ou inférieur à 0.

Par conséquent, d_s reflète l'importante des différences observées. Plus les taux observés seront proches des probabilités de réponses sous l'hypothèse nulle correspondante, plus d_{is} sera petit et plus leur somme, d_s sera petite.

S_s quant à lui, permet de déterminer le sens des différences :

- si les taux observés dans les sous-populations i sont supérieurs strictement aux π_{0i} respectifs, alors $S_s = 2$
- si les taux observés dans les sous-populations i sont inférieurs strictement aux π_{0i} respectifs, alors $S_s = -2$
- si le taux observé dans une sous-population i est supérieur strictement à son π_{0i} et si le taux observé dans l'autre sous-population i est inférieur strictement à son π_{0i}, alors $S_s = 0$
- si les taux observés dans les sous-populations i sont strictement égaux aux π_{0i}, $S_s = 0$
- si le taux observé est strictement égal à π_{0i} dans une seule des deux sous-populations, $S_s = -1$ ou 1 selon le sens de la différence dans l'autre sous-population.

Le test d'hétérogénéité de niveau γ sous H_{00} a la forme suivante à l'étape s :

$$T_s = Ind\{d_s > c_s \; et \; S_s = 0\}$$

où $Ind\{x\} = 1$ si la condition x est vérifiée et 0 sinon.

c_s est une fonction de γ_B :

$$c_s = \min(c|P(d_s > c \; et \; S_s = 0|H_{00}) \leq \gamma_B).$$

- En l'absence d'hétérogénéité, $T_s = 0$.
- En cas d'hétérogénéité, $T_s = 1$.

Avec la méthode B, on définit finalement Ψ_s de la façon suivante :

- $\Psi_s = (Ind\{d_{1s} > 0\} + 2 \times Ind\{d_{2s} > 0\}) \times T_s$.

On retrouve bien que :

- lorsqu'il n'y a pas d'hétérogénéité, $\Psi_s = 0$
- lorsqu'il y a une hétérogénéité en faveur de la sous-population 1, alors $\Psi_s = 1$
- lorsqu'il y a une hétérogénéité en faveur de la sous-population 2, alors $\Psi_s = 2$

La figure 4 ci-dessous synthétise la différence de concept entre les méthodes A et B pour la définition de l'hétérogénéité.

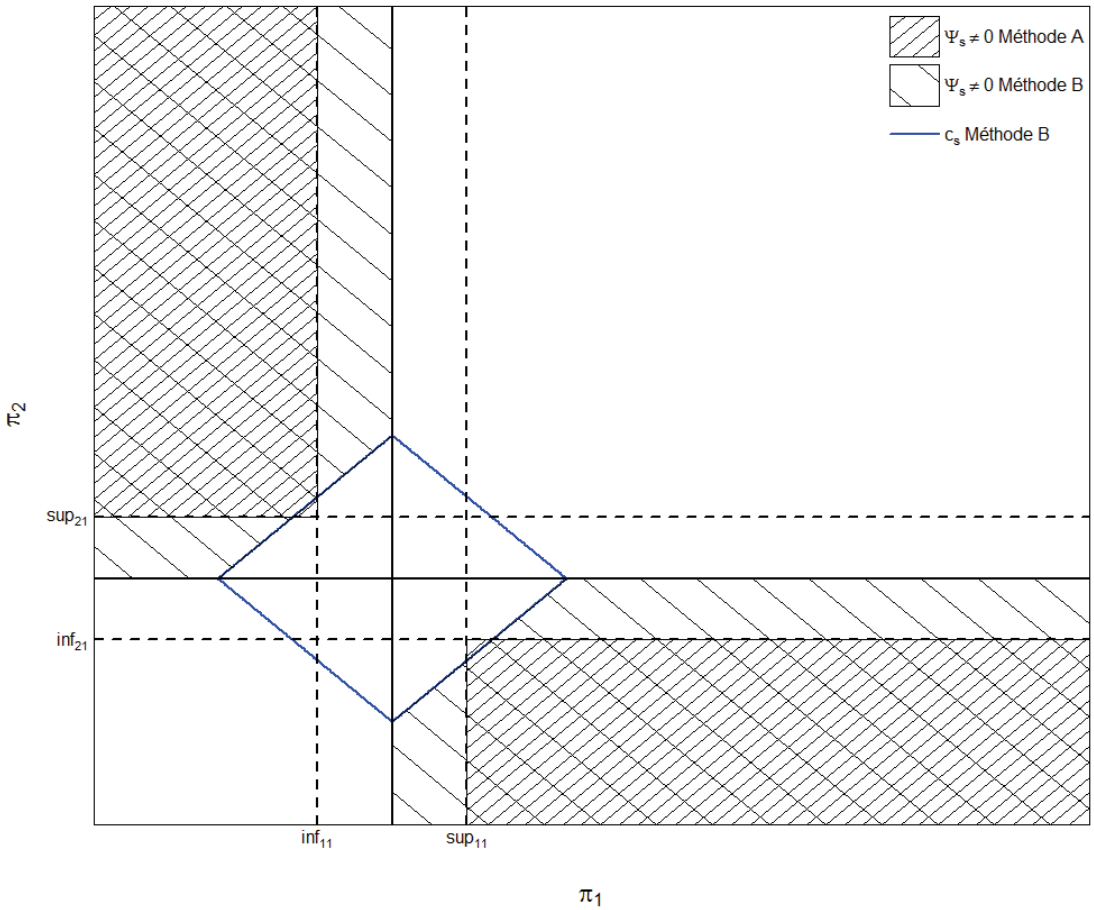

Figure 4. Différence de concept entre les méthodes A et B pour la définition de l'hétérogénéité.

Attention : le γ_B de la méthode B correspond formellement à $\frac{\gamma^2}{2}$ de la méthode A.

EVALUATION DES CARACTERISTIQUES OPERATOIRES

a) Calculs théoriques

Comme dans le chapitre précédent, tous les résultats sont obtenus en réalisant des calculs exacts en utilisant la loi binomiale.

Nous avons comparé la méthode B au schéma de Fleming classique (Méthode H) et à la méthode A.

Nous avons étudié des probabilités de succès comprises entre 0 et 1, par pas de 0.01 pour déterminer, comme précédemment :

- L'effectif maximal à inclure
- Le nombre attendu de patients sous les hypothèses nulles et alternatives $E(n|H_{00})$, $E(n|H_{01})$, $E(n|H_{10})$, $E(n|H_{11})$
- Le taux de bonnes conclusions
- La probabilité de conclure à l'Inefficacité ou à l'Efficacité pour l'ensemble de la population alors que l'on est sous les hypothèses H_{01} ou H_{10},
- La probabilité de détecter une sous-population non sensible à la nouvelle thérapeutique dès l'étape 1 sous H_{01} ou H_{10}
- Et la probabilité de continuer en phase III sous H_{00}.

Les mêmes exemples qu'au chapitre 2 sont rapportés :

$\pi_{01} = \pi_{02}$, $\Delta_1 = \Delta_2$, $\omega = 1$, $\alpha = 0.05$, $\beta = 0.1$ et $\gamma_B = \frac{0.6^2}{2} = 0.18$ et pour $\pi_{0i} = 0.25$, $\gamma_B = 0.045$ *et* 0.32. .

b) Application du nouveau plan aux données de l'essai REMAGUS 02

On rappelle les hypothèses nulles et alternatives des deux plans de Fleming conduits en parallèle dans l'essai REMAGUS 02 :

- Sous-population des femmes HER_{2+} :
  H_0 : $\pi_1 \leq 0.15$ & H_1 : $\pi_1 > 0.15$ (avec un taux d'efficacité optimale de 0.30), α=0.05 et β=0.10
- Sous-population des femmes HER_{2-} :
  H_0 : $\pi_2 \leq 0.15$ & H_1 : $\pi_2 > 0.15$ (avec un taux d'efficacité optimale de 0.25), α=0.05 and β=0.10

Comme pour la méthode A, la prévalence des femmes HER_{2+} est fixée à 25% (soit $\omega = 3$) et $\gamma_B = \frac{0.6^2}{2} = 0.18$.

Les conclusions finales et les effectifs inclus sont comparés entre :

- La nouvelle version du schéma adaptatif (« Méthode B »),
- Un plan de Fleming classique (« Méthode H ») : dans ce travail nous avons tenu compte de la prévalence attendue de la positivité des récepteurs, ici fixée à 25%.
- L'essai REMAGUS 02 tel qu'il a été conduit, c'est-à-dire en mettant en œuvre parallèlement deux plans de Fleming (« Méthode D »), mais incluant dans les résultats les réponses simulées pour les patients supplémentaires.

RESULTATS THEORIQUES ET APPLICATION AUX DONNEES REELLES

RESULTATS THEORIQUES

a) Adéquation des méthodes A, B et H en termes d'effectifs (Tableau 23)

Les effectifs attendus sont toujours très similaires entre les méthodes A, B et H en particulier quand les taux de réponses sont proches des hypothèses nulles $E(n|H_{00})$, ou alternatives $E(n|H_{11})$. Sous H_{01} ou H_{10}, jusqu'à 20% de patients supplémentaires sont susceptibles d'être inclus (contre 10% avec la méthode A) mais on rappelle que ce sont ceux qui sont sensibles à la nouvelle thérapeutique.

Tableau 23. Effectifs maximaux et attendus avec les méthodes A, B et H sous les hypothèses nulle, alternative, ou combinée pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, w=1, α=0.05 et β=0.1

π_{0i}	$\gamma^{\dagger}$	Effectif maximal			Effectif attendu								
					A			B			H		
		A	B	H	$H_{00}^{\S}$	H_{01} ou H_{10}	H_{11}	$H_{00}^{\S}$	H_{01} ou H_{10}	H_{11}	H_{00}	H_{01} ou H_{10}	H_{11}
0.05	**0.6**	32	32	28	21.11	24.77	21.05	21.36	26.14	21.57	21.11	24.77	21.05
0.10	**0.6**	42	42	36	27.78	32.09	27.36	28.83	33.96	27.74	27.78	32.09	27.36
0.15	**0.6**	48	48	40	32.35	37.56	31.93	32.63	38.87	32.47	31.78	36.39	31.78
0.20	**0.6**	57	57	48	38.11	44.23	35.70	37.72	44.53	35.96	36.66	41.95	35.42
0.25	**0.3**	62	62	52	38.74	47.29	41.86	39.08	48.60	42.22	38.46	46.43	41.80
0.25	**0.6**	62	62	52	39.88	48.23	41.93	40.70	50.40	42.63	38.46	46.43	41.80
0.25	**0.8**	62	62	52	40.73	49.85	42.27	43.05	51.95	43.05	38.46	46.43	41.80
0.30	**0.6**	67	67	56	43.09	51.84	43.78	44.38	54.32	44.58	41.01	49.27	43.59
0.35	**0.6**	69	69	56	41.34	52.49	45.98	43.39	54.64	46.48	38.62	48.33	45.59
0.40	**0.6**	70	70	56	42.25	52.59	44.63	44.93	55.02	45.23	40.34	48.86	44.24
0.45	**0.6**	68	68	56	41.26	52.02	46.67	42.95	54.71	47.48	38.08	48.05	46.41
0.50	**0.6**	67	67	56	42.51	52.75	45.46	44.16	54.41	45.78	39.73	48.96	45.20
0.55	**0.6**	67	67	56	40.26	52.99	48.12	42.80	54.96	48.47	37.46	48.02	47.79
0.60	**0.6**	60	60	48	34.25	44.39	40.97	35.59	46.63	41.54	31.79	40.57	40.80
0.65	**0.6**	53	53	44	30.43	40.42	40.29	31.37	41.85	40.57	28.63	37.18	40.18
0.70	**0.6**	47	47	40	25.94	35.41	36.77	29.23	38.83	37.08	24.73	32.42	36.70
0.75	**0.6**	38	38	32	19.66	26.59	31.32	21.91	30.33	31.49	19.15	24.93	31.31

§Sous H_{00}, $\pi_1 = \pi_{01}$ & $\pi_2 = \pi_{02}$; sous H_{01}, $\pi_1 = \pi_{01}$ & $\pi_1 = \pi_{12}$; sous H_{10}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{02}$; sous H_{11}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{12}$; † $\gamma_B = \frac{\gamma^2}{2}$

b) Amélioration de la pertinence des décisions avec la méthode B

Augmentation du taux de bonnes conclusions finales (Tableau 24)

Sous H_{01} ou H_{10}, le taux de bonnes conclusions conduisant à des essais de phase III sur la seule sous-population efficace, varie de 23.7% à 42.2% (contre 26.3% précédemment). En revanche, du fait dans certains cas d'une détection à tort de l'hétérogénéité, les taux de bonnes conclusions sous H_{00} et H_{11} sont légèrement inférieurs avec la méthode B comparativement aux méthodes A et H.

Tableau 24. ***Taux de bonnes conclusions finales des méthodes A, B et H sous les hypothèses nulle H_{00}, alternative H_{11}, ou combinée H_{01} et H_{10} pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, ω=1, α=0.05 et β=0.1***

π_{0i}	$\gamma^{\dagger}$	H_{00} [§]			H_{01}/H_{10}			H_{11}		
		A	**B**	**H**	**A**	**B**	**H**	**A**	**B**	**H**
0.05	**0.6**	0.953	0.948	0.953	0.000	0.415	0	0.936	0.795	0.936
0.10	**0.6**	0.974	0.955	0.974	0.018	0.324	0	0.886	0.841	0.887
0.15	**0.6**	0.958	0.952	0.968	0.145	0.316	0	0.866	0.806	0.878
0.20	**0.6**	0.937	0.941	0.952	0.217	0.286	0	0.896	0.858	0.917
0.25	**0.3**	0.953	0.950	0.957	0.072	0.243	0	0.906	0.864	0.910
0.25	**0.6**	0.950	0.944	0.957	0.104	0.333	0	0.906	0.839	0.910
0.25	**0.8**	0.941	0.935	0.957	**0.263**	**0.422**	0	0.881	0.812	0.910
0.30	**0.6**	0.939	0.932	0.948	0.135	0.365	0	0.919	0.847	0.924
0.35	**0.6**	0.939	0.934	0.951	0.163	0.311	0	0.901	0.870	0.910
0.40	**0.6**	0.938	0.932	0.950	0.195	0.346	0	0.895	0.855	0.911
0.45	**0.6**	0.946	0.938	0.956	0.135	0.347	0	0.898	0.840	0.902
0.50	**0.6**	0.946	0.939	0.958	0.167	0.289	0	0.902	0.880	0.908
0.55	**0.6**	0.956	0.950	0.966	0.186	0.325	0	0.898	0.872	0.905
0.60	**0.6**	0.949	0.943	0.958	0.122	0.305	0	0.902	0.860	0.903
0.65	**0.6**	0.966	0.960	0.973	0.110	0.256	0	0.876	0.851	0.877
0.70	**0.6**	0.970	0.957	0.977	0.107	0.275	0	0.887	0.882	0.887
0.75	**0.6**	0.973	0.963	0.977	0.074	**0.237**	0	0.910	0.910	0.910

[§] *Sous H_{00}, $\pi_1 = \pi_{01}$ & $\pi_2 = \pi_{02}$; sous H_{01}, $\pi_1 = \pi_{01}$ & $\pi_1 = \pi_{12}$; sous H_{10}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{02}$; sous H_{11}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{12}$;* [†] $\gamma_B = \frac{\gamma^2}{2}$

Diminution du taux d'arrêt complet du développement pour inefficacité pour des sous-populations sensibles (Tableau 25)

Les probabilités de conclure à l'inefficacité pour l'ensemble de la population avec la méthode B sont bien inférieures à celles des méthodes A et H sous H_{01} ou $H_{10.}$

Tableau 25. ***Probabilité de conclure à l'Inefficacité pour la population entière avec les méthodes A, B et H sous les combinée H_{01} et H_{10} pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, ω=1, α=0.05 et β=0.1***

π_{0i}	$\gamma^{\dagger}$	H_{01}/H_{10}		
		A	**B**	**H**
0.05	**0.6**	0.381	0.306	0.381
0.10	**0.6**	0.562	0.373	0.562
0.15	**0.6**	0.482	0.408	0.582
0.20	**0.6**	0.381	0.391	0.516
0.25	**0.3**	0.495	0.451	0.545
0.25	**0.6**	0.472	0.383	0.545
0.25	**0.8**	0.381	0.315	0.545
0.30	**0.6**	0.427	0.341	0.519
0.35	**0.6**	0.428	0.367	0.549
0.40	**0.6**	0.423	0.361	0.558
0.45	**0.6**	0.476	0.373	0.588
0.50	**0.6**	0.465	0.394	0.598
0.55	**0.6**	0.472	0.400	0.629
0.60	**0.6**	0.510	0.413	0.622
0.65	**0.6**	0.598	0.504	0.706
0.70	**0.6**	0.627	0.472	0.734
0.75	**0.6**	0.681	0.518	0.755

[§]*Sous* H_{01}, $\pi_1 = \pi_{01}$ & $\pi_1 = \pi_{12}$; *sous* H_{10}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{02}$;[†] $\gamma_B = \frac{\gamma^2}{2}$

Diminution du taux de poursuite du développement pour efficacité pour des sous-populations non sensibles (Tableau 26)

De même, les probabilités sous H_{01}/H_{10} de conclure à l'Efficacité pour la population entière avec les méthodes A et H sont bien supérieures à celles de la méthode B.

Tableau 26. ***Probabilité de conclure à l'Efficacité pour la population entière avec les méthodes A, B et H sous les combinée H_{01} et H_{10} pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, ω=1, α=0.05 et β=0.1***

π_{0i}	$\gamma^{\dagger}$	H_{01}/H_{10}		
		A	**B**	**H**
0.05	**0.6**	0.619	0.277	0.619
0.10	**0.6**	0.420	0.301	0.438
0.15	**0.6**	0.373	0.275	0.418
0.20	**0.6**	0.401	0.322	0.484
0.25	**0.3**	0.433	0.306	0.455
0.25	**0.6**	0.424	0.283	0.455
0.25	**0.8**	0.355	0.261	0.455
0.30	**0.6**	0.438	0.292	0.481
0.35	**0.6**	0.408	0.321	0.451
0.40	**0.6**	0.382	0.292	0.442
0.45	**0.6**	0.389	0.278	0.412
0.50	**0.6**	0.368	0.316	0.402
0.55	**0.6**	0.341	0.274	0.371
0.60	**0.6**	0.368	0.278	0.378
0.65	**0.6**	0.292	0.240	0.294
0.70	**0.6**	0.266	0.253	0.266
0.75	**0.6**	0.245	0.245	0.245

[§]*Sous* H_{01}, $\pi_1 = \pi_{01}$ & $\pi_1 = \pi_{12}$; *sous* H_{10}, $\pi_1 = \pi_{11}$ & $\pi_1 = \pi_{02}$;[†] $\gamma_B = \frac{\gamma^2}{2}$

Détection non négligeable de l'hétérogénéité dès la première étape (Tableau 27)

La probabilité de détecter une hétérogénéité de réponses est nettement supérieure avec la méthode B par rapport à la méthode A.

Comme pour le taux de bonnes conclusions, l'identification d'une hétérogénéité à tort sous H_{00} ou H_{11}, conduit à la poursuite ou à l'abandon à tort à la fin de l'étape 1 de l'une ou l'autre des deux sous-populations.

Tableau 27. ***Détection de l'hétérogénéité à l'étape 1 avec la méthode A sous les hypothèses nulle H_{00}, alternative H_{11}, ou combinée H_{01} et H_{10} pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, ω=1, α=0.05 et β=0.1***

π_{0i}	$\gamma^{\dagger}$	A			B		
		$H_{00}^{\S}$	H_{01}/H_{10}	H_{11}	$H_{00}^{\S}$	H_{01}/H_{10}	H_{11}
0.05	**0.6**	0	0	0	0.062	0.394	0.148
0.10	**0.6**	0	0	0	0.174	0.321	0.065
0.15	**0.6**	0.071	0.148	0.020	0.106	0.320	0.090
0.20	**0.6**	0.113	0.217	0.030	0.069	0.267	0.066
0.25	**0.3**	0.020	0.073	0.006	0.039	0.197	0.045
0.25	**0.6**	0.052	0.099	0.008	0.107	0.313	0.080
0.25	**0.8**	0.137	0.262	0.042	0.231	0.430	0.119
0.30	**0.6**	0.071	0.128	0.010	0.140	0.346	0.086
0.35	**0.6**	0.081	0.166	0.017	0.150	0.292	0.052
0.40	**0.6**	0.084	0.196	0.024	0.164	0.323	0.065
0.45	**0.6**	0.086	0.138	0.010	0.169	0.347	0.077
0.50	**0.6**	0.090	0.167	0.013	0.156	0.288	0.040
0.55	**0.6**	0.086	0.194	0.015	0.169	0.318	0.045
0.60	**0.6**	0.071	0.127	0.006	0.134	0.298	0.053
0.65	**0.6**	0.060	0.116	0.004	0.109	0.256	0.035
0.70	**0.6**	0.045	0.111	0.002	0.175	0.289	0.019
0.75	**0.6**	0.023	0.076	0	0.125	0.246	0.008

§*Sous* H_{00}, $\pi_1=\pi_{01}$ & $\pi_2=\pi_{02}$; *sous* H_{01}, $\pi_1=\pi_{01}$ & $\pi_1=\pi_{12}$; *sous* H_{10}, $\pi_1=\pi_{11}$ & $\pi_1=\pi_{02}$; *sous* H_{11}, $\pi_1=\pi_{11}$ & $\pi_1=\pi_{12}$;† $\gamma_B=\frac{\gamma^2}{2}$

Contrôle satisfaisant des risques d'erreur (Tableau 28)

La probabilité de passer en phase III sous H_{11} est encore plus élevée avec la méthode B par rapport aux méthodes A et H et le risque de première espèce est également bien contrôlé avec cette nouvelle méthode (bien qu'un peu plus élevé qu'avec la méthode B).

Tableau 28. ***Probabilité de passer en phase III avec les méthodes A et H sous les hypothèses nulle H_{00}, ou alternative H_{11}, pour des taux de réponse identiques sous l'hypothèse nulle, Δ_i=0.2, ω=1, α=0.05 et β=0.1***

π_{0i}	$\gamma^{\dagger}$	H_{00} [§]			H_{11}		
		A	**B**	**H**	**A**	**B**	**H**
0.05	**0.6**	0.047	0.052	0.047	0.936	0.942	0.936
0.10	**0.6**	0.026	0.045	0.026	0.887	0.905	0.887
0.15	**0.6**	0.042	0.048	0.032	0.885	0.895	0.878
0.20	**0.6**	0.063	0.059	0.048	0.926	0.925	0.917
0.25	**0.3**	0.047	0.050	0.043	0.912	0.915	0.910
0.25	**0.6**	0.050	0.056	0.043	0.914	0.921	0.910
0.25	**0.8**	0.059	0.065	0.043	0.921	0.929	0.910
0.30	**0.6**	0.061	0.068	0.052	0.929	0.935	0.924
0.35	**0.6**	0.061	0.066	0.049	0.918	0.923	0.910
0.40	**0.6**	0.062	0.068	0.050	0.919	0.924	0.911
0.45	**0.6**	0.054	0.062	0.044	0.908	0.916	0.902
0.50	**0.6**	0.054	0.061	0.042	0.914	0.920	0.908
0.55	**0.6**	0.044	0.050	0.034	0.913	0.918	0.905
0.60	**0.6**	0.051	0.057	0.042	0.908	0.914	0.903
0.65	**0.6**	0.034	0.040	0.027	0.880	0.886	0.877
0.70	**0.6**	0.030	0.043	0.023	0.890	0.901	0.887
0.75	**0.6**	0.027	0.037	0.023	0.911	0.918	0.910

[§] *Sous H_{00}, $\pi_1=\pi_{01}$ & $\pi_2=\pi_{02}$; sous H_{01}, $\pi_1=\pi_{01}$ & $\pi_1=\pi_{12}$; sous H_{10}, $\pi_1=\pi_{11}$ & $\pi_1=\pi_{02}$; sous H_{11}, $\pi_1=\pi_{11}$ & $\pi_1=\pi_{12}$;* [†] $\gamma_B=\frac{\gamma^2}{2}$

APPLICATION PRATIQUE

Les résultats issus de l'application aux données réelles REMAGUS 02 sont présentés dans le Tableau 29.

Avec la méthode que nous proposons, seules 106 femmes auraient participé à l'étude : 64 femmes HER_{2+} et 42 femmes HER_{2-}.

Ici plus de femmes HER_{2+} que de femmes HER_{2-} participent à l'étude car à la fin de l'étape 1 seule la sous-population des femmes HER_{2+} est jugée comme sensible aux nouvelles thérapeutiques. Une hétérogénéité de réponses est détectée dès la fin de l'étape 1 ($T_1 = 1$). En effet, on a d'une part d_1=0.24 supérieur à c_1=0.15 ; on a d'autre part $S_1 = 0$ puisque le taux de réponses observé dans la sous-population des femmes HER_{2-} (12%) est inférieur au taux de réponse sous l'hypothèse nulle (15%) alors qu'il est de 36% dans la sous-population des femmes HER_{2+}, bien supérieur au taux de réponse sous l'hypothèse nulle (15%).

La conclusion de cet essai aurait été : Efficacité des nouvelles thérapeutiques pour les femmes HER_{2+} et Inefficacité pour les femmes HER_{2-}. Du fait de la poursuite de l'étude en étape 2 par les femmes HER_{2+} uniquement, autant de femmes HER_{2+} auraient participé à la recherche que dans l'essai REMAGUS 02 mais par contre ce sont seulement 42 femmes HER_{2-} soit 94 de moins que dans l'essai REMAGUS 02 qui auraient été exposées inutilement aux traitements expérimentaux.

Ce nouveau plan de phase II, adaptatif et stratifié, aurait donc conduit à la même conclusion que l'essai REMAGUS 02 réellement conduit avec beaucoup moins de patientes, en particulier celles qui ne sont pas sensibles aux nouvelles thérapeutiques car l'inclusion de ce sous-groupe de femmes aurait été interrompue dès la fin de la première étape.

Tableau 29. Hypothèses, règles de décision, résultats observés et conclusions finales en fonction du schéma d'étude suivi

		Méthode A		Méthode H	Méthode B		Méthode D	
Sous-Population*		1 : 25%	2 : 75%	1 : 25% 2 : 75%	1 : 25%	2 : 75%	1 : 100%	2 : 100%
Hypothèses		π_0 : 0.15 π_1 : 0.30	π_0 : 0.15 π_1 : 0.25	π_0 : 0.15 π_1 : 0.26	π_0 : 0.15 π_1 : 0.30	π_0 : 0.15 π_1 : 0.25	π_0 : 0.15 π_1 : 0.30	π_0 : 0.15 π_1 : 0.25
		α : 0.05 β : 0.10 γ : 0.6		α : 0.05 β : 0.10	α : 0.05 β : 0.10 γ_B : 0.18		α : 0.05 β : 0.10	α : 0.05 β : 0.10
Effectifs par étape	**Etape 1**	14	42	56 (14+42)	14	42	32	68
	Etape 2	14 ou 50	42 ou 94	56 (14+42)	14 ou 50	42 ou 94	32	68
Etape 1	**Réponses observées**	$\Omega_1 = 0$		$\Omega_1 = 0$	$\Omega_1 = 0$		$\Omega_1 = 0$	$\Omega_1 = 0$
	Détection de l'hétérogénéité	Ψ_1=0			Ψ_1=1			
	Décision	C_1C_2		C_1C_2	C_1I_2		C_1	C_2
Etape 2	**Réponses observées**	$\Omega_2 = 0$		$\Omega_2 = 0$	$\Omega_2 = 1$		$\Omega_2 = 1$	$\Omega_2 = 0$
Conclusion à la fin de l'essai		I_1I_2		I_1I_2	E_1I_2		E_1	I_2
Nombre de participantes à l'essai		42+14+42+14 =**112**		56+56=**112**	42+14+50=**106**		32+32+68+68 =**200**	

**1= Femmes HER_2+, 2= Femmes HER_2-*
I=Inefficacité (de la sous-population 1, I_1 ou 2, I_2) ; E=Efficacité (de la sous-population 1, E_1 ou 2, E_2) ; C=Poursuite à l'étape 2 (avec la sous-population 1, E_1 ou 2, E_2).

SYNTHESE

La nouvelle version de la méthode que nous proposons dans cette partie permet effectivement d'améliorer la détection de l'hétérogénéité quand elle existe entre deux sous-populations.

Ainsi le taux de bonnes conclusions finales et la probabilité de détecter la sous-population non sensible dès l'étape 1 sous H_{01} ou H_{10} ont doublé, les probabilités de conclure à l'inefficacité ou à l'efficacité pour la population entière, également sous H_{01} ou H_{10}, ont diminué de près de 15%, tout en contrôlant le risque de première espèce de façon tout à fait satisfaisante.

Appliquée aux données réelles de l'essai REMAGUS 02, cette nouvelle méthodologie des essais de phase II permettant de tenir compte de l'hétérogénéité de la population participant à l'essai par l'identification initiale de deux strates a pu montrer tout son intérêt pratique en recherche clinique.

DISCUSSION GENERALE

Il existe une importante variabilité clinique et biologique des individus et de leurs tumeurs. Faute de moyens et de connaissances, les chimiothérapies ont longtemps été prescrites sans connaissance fine du patient et de son cancer, et ce, tant que l'apport de ces thérapeutiques en termes d'efficacité était important. Avec une diminution de cet apport pour les nouvelles thérapeutiques proposées selon les modèles classiques, il est devenu nécessaire de développer de nouvelles stratégies thérapeutiques et d'identifier de nouvelles cibles potentielles. Ces dernières années la connaissance des caractéristiques individuelles est devenue l'un des éléments-clés de la réponse à la thérapie. Les progrès de la recherche fondamentale ont en effet permis d'établir que la cellule tumorale interagissait avec son environnement. La recherche académique et industrielle a développé des molécules dirigées contre ces cibles thérapeutiques. Les thérapeutiques ciblées, par un mécanisme non directement cytotoxique, visent à contrôler la maladie sur une longue période. Selon leur nature et leur mode d'action, ces molécules vont s'intégrer dans une stratégie thérapeutique globale où elles feront partie de schémas utilisant conjointement la chimiothérapie et/ou l'hormonothérapie et/ou la radiothérapie. De ce fait, le nombre de schémas thérapeutiques possibles devient élevé et le profil biologique de chaque tumeur oriente la décision thérapeutique vers un traitement de plus en plus personnalisé.

L'importance de la prise en compte des biomarqueurs a été évoquée dès 1989 par Blackledge qui conclut dans l'un de ses articles (Blackledge, 1989) que les taux de réponse observés en phase II peuvent dépendre non seulement de l'efficacité de la thérapeutique mais également de la proportion de patients inclus qui ont une très faible probabilité de réponse. Ne pas en tenir compte pourrait donc se traduire par des essais de phase II négatifs car incapables de détecter l'efficacité ciblée de nouvelles thérapeutiques.

Cependant, la prise en compte des biomarqueurs dès la planification des essais de phase II soulève des questions méthodologiques (sur leur mesure, leur incorporation aux méthodes statistiques et leur interprétation) ; le

développement des méthodes séquentielles et adaptatives entre dans une logique d'amélioration des méthodes classiques et est largement encouragé par les experts internationaux (Cannistra, 2009 ; Seymour, 2010). L'idée n'est donc plus seulement de se concentrer uniquement sur les questions cliniques les plus pertinentes mais également d'améliorer la méthodologie des essais de phase II pour réduire le nombre d'échecs en phase III.
En effet, approximativement 60% des thérapies ayant une activité prometteuse lors d'essais de phase II à un seul bras n'arrivent pas à démontrer leur supériorité lors des essais de phase III (Kola, 2004). Si ce pourcentage de faux-positifs pouvait être acceptable dans les années 80 parce qu'il n'y avait pas beaucoup de thérapeutiques efficaces et innovantes, les choses ont beaucoup changé ces dernières années, en particulier depuis l'arrivée des thérapies ciblées.

Le coût des essais de phase III et des traitements anti-cancéreux d'une manière générale, les questions éthiques qui se posent devant l'exposition de patients à des thérapeutiques potentiellement toxiques et l'investissement demandé aux investigateurs pour inclure et suivre les participants ont conduit à une réflexion sur la rigueur avec laquelle sont conduits les essais de phase II.
Malheureusement, alors que la prise en charge thérapeutique des patients atteints de cancer vise à une approche plus personnalisée en proposant des thérapeutiques mieux ciblées, peu de méthodologies d'essais de phase II ont été développées pour prendre en considération les caractéristiques de la population dès la planification de l'essai.

La méthodologie que nous proposons à l'issue de ce travail s'inscrit parfaitement dans cette tendance puisqu'il s'agit d'une méthodologie de phase II

- visant à réduire les échecs de développement du médicament en phase II et III
- intégrant un paramètre d'identification de sous-populations (ce paramètre pouvant être un biomarqueur ou une caractéristique autre

telle que l'historique des traitements reçus par exemple) et servant à modifier l'inclusion des participants s'il apparaît que ce biomarqueur peut être un élément lié à la réponse au traitement.

Ainsi, la population cible, c'est-à-dire celle pouvant à terme bénéficier de la nouvelle thérapeutique, est mieux définie.

Dans cette thèse, nous proposons deux méthodes (A et B) qui sont des extensions de la méthode de Fleming à 2 étapes. Pour chaque méthode, les participants sont scindés en deux-populations en fonction d'une caractéristique particulière. Le principe des deux méthodes est de s'intéresser au taux de réponses global et aux taux de réponses dans chacune des sous-populations pour éventuellement mettre en évidence une hétérogénéité de réponses au traitement entre les deux sous-populations. La seconde méthode (Méthode B) est une amélioration de la première (Méthode A) en termes d'identification de l'hétérogénéité.

Cette nouvelle méthodologie présente plusieurs atouts.

- Tout d'abord elle ne se contente pas seulement de faire l'hypothèse d'une hétérogénéité des participants vis-à-vis de la réponse au traitement et d'en tenir compte dans la statistique de test pour l'évaluation de l'efficacité de la nouvelle thérapeutique. Elle permet d'identifier si l'une des deux sous-populations est effectivement sensible au traitement dans le cas où la population entière participant à l'étude serait hétérogène.

- Elle est une extension de la méthode de Fleming qui permet l'arrêt précoce pour efficacité ou inefficacité. Cette méthode est classiquement utilisée. Ainsi les paramètres à définir avec l'investigateur sont connus et habituellement employés, en dehors de

la probabilité d'identifier à tort une hétérogénéité de réponses entre les deux sous-populations sous H_{00}.

- La conduite de l'essai est simple, le nombre de sujets maximal pouvant être inclus est connu dès la rédaction du protocole, ce qui permet de le comparer aux possibilités de recrutement.

- Le risque de première espèce est bien maîtrisé, permettant de limiter le nombre de faux positifs à l'issu de la phase II

- Le taux de bonnes conclusions sous H_{01} ou H_{10} (c'est-à-dire sous l'hypothèse où l'une des deux sous-populations est sensible à la nouvelle thérapeutique) est d'un peu plus de 30%. Si ce chiffre n'est pas aussi satisfaisant que ce que l'on aurait pu espérer, il signifie tout de même que l'on diminue de 30% le nombre d'essais de phase II qui sont soit dits négatifs alors qu'une partie de la population pourrait bénéficier de la thérapeutique (arrêt à tort en phase II du développement du médicament), soit dits positifs alors qu'une partie de la population n'en bénéficie pas (entraînant non seulement le risque de rendre l'essai de phase III négatif, à tort puisqu'une partie de la population est en fait sensible à la thérapeutique, mais également d'exposer inutilement un grand nombre de patients non sensibles à une thérapeutique toxique).

- Enfin, les effectifs attendus avec notre méthode sont similaires à ceux d'un plan de Fleming classique en cas d'efficacité ou d'inefficacité pour l'ensemble de la population. En cas d'hétérogénéité, quelques patients supplémentaires de la sous-population d'intérêt sont inclus.

Certains points sont critiquables.

- La caractéristique permettant de scinder la population en deux doit être connue avant le début de l'essai. Des recherches sur les

mécanismes d'action de la nouvelle thérapeutique pouvant expliquer le rôle de la caractéristique sur la réponse au traitement doivent par conséquent avoir été menés préalablement. La méthodologie ne permet pas de tenir compte d'un nouveau facteur identifié seulement en cours d'essai. Cette caractérisation éventuellement moléculaire préalable de la tumeur a par ailleurs un coût. Cependant des études de coût ont été menées, comme par exemple dans le cas du cancer du sein et de la recherche de la mutation HER_2 (Elkin, 2004). Et il a pu être démontré que la détermination du statut HER_2, outre l'intérêt clinique, est toujours plus coût-efficace que l'administration de la chimiothérapie (en l'occurrence du Trastuzumab) à toutes les patientes, quel que soit le protocole utilisé pour rechercher le statut HER_2.

- Le ratio entre les deux sous-populations est fixé au début de l'essai et constant entre les deux étapes. Les participants à l'essai sont donc recrutés à la fois sur leur maladie et sur cette caractéristique, dans la limite des effectifs planifiés par sous-population. Ce critère de sélection supplémentaire peut alors devenir un handicap pour l'investigateur si l'hypothèse faite sur la prévalence de ce critère s'avère ne pas être tout à fait juste.

- L'essai n'est pas randomisé. Le rôle prédictif plutôt que pronostique de la caractéristique ne peut donc pas être clairement établi. Malgré cela, l'intérêt de cette distinction n'est pas évident à ce stade de l'évaluation de la thérapeutique. En effet, l'objectif de cette nouvelle approche est d'éviter la dilution de l'effet traitement en phase II ou III et donc l'abandon à tort de thérapeutiques intéressantes pour certains patients. Si suite à la phase II, l'évaluation est poursuivie pour une partie seulement de la population, que cette population ait été sélectionnée parce qu'elle présentait un meilleur pronostic ou parce que du fait de sa caractéristique la réponse au traitement était meilleure, l'objectif de ciblage thérapeutique aura été atteint : la méthode aura permis de mieux cibler la population évoluant

favorablement sous traitement. Une restriction des critères d'inclusion en phase III pourra être proposée. Elle évitera que la participation en grand nombre de sujets à l'évolution défavorable ne nuise à l'identification d'une thérapeutique intéressante pour certains malades. Elle évitera également de prescrire un traitement inutile et potentiellement toxique aux patients non sensibles à la nouvelle thérapeutique.

- Enfin, la méthode B permet d'identifier une hétérogénéité de réponse dans un plus grand nombre de cas par rapport à la méthode A mais le risque de ne pas poursuivre l'évaluation avec une sous-population sensible au traitement augmente avec le risque γ dont la définition, nouvelle pour l'investigateur qui a l'habitude d'élaborer des essais de phase II selon les plans de Fleming ou Simon, n'est pas toujours facile à appréhender (comme cela peut être déjà le cas des risques α et β).

Le travail réalisé au cours de cette thèse ouvre d'autres perspectives de travail et d'amélioration.

Des suites à ce travail sont déjà en cours de réalisation comme par exemple l'application du principe de notre méthode à un plan en une seule étape ou à un plan de Simon. Il est intéressant d'étendre notre méthode à ce dernier schéma car il est l'un des plus utilisés par les cliniciens (Thezenas, 2004). Un article, intitulé « A new Simon based adaptative design for phase II trials allowing detection of heterogeneity » sera bientôt soumis à la revue Clinical Trials.

On peut également souligner qu'un essai thérapeutique de phase II, de promotion industrielle, est actuellement en cours d'élaboration suivant la méthode que nous proposons.

Il serait encore intéressant :

- d'augmenter le nombre d'étapes pour permettre d'augmenter plus progressivement le nombre de sujets inclus dans l'essai (et donc de réduire potentiellement le nombre de sujets inclus à tort)

- d'augmenter le nombre sous-populations étudiées par l'augmentation du nombre de caractéristiques prises en compte, et ce pour affiner encore les profils de patients sensibles (mais les décisions possibles deviennent très nombreuses et le critère d'hétérogénéité doit être complètement repensé)

- d'étudier l'impact sur la probabilité de passer en phase III à tort ou à raison d'une modification des conclusions possibles à l'étape 2, notamment quand le nombre cumulé de réponses est faible et qu'une hétérogénéité de réponses est détectée,

- d'intégrer le risque γ dans le calcul des effectifs et des bornes d'arrêt (comme α et β actuellement) pour prendre en compte l'ensemble des risques,

- de travailler sur un bi-critère en intégrant les données de toxicité pour réduire les risques d'exposition inutiles aux thérapeutiques toxiques,

- de prendre en compte d'autres critères que la réponse en tant que critère binaire, comme le délai de survie par exemple, plus largement utilisé en phase III en cancérologie (le problème étant alors celui du délai d'obtention de la mesure du critère de jugement pour un essai de phase II),

- de réfléchir à la question de la randomisation et à son impact en termes de règles de décision.

En conclusion, si le nombre de gènes et de fonctions cellulaires qui favorisent la transformation d'une cellule normale en cellule tumorale fait que le nombre de cibles thérapeutiques potentielles est extrêmement important, il faut trouver des cibles suffisamment spécifiques de la cellule cancéreuse pour la rendre repérable sans ambigüité et suffisamment importantes dans le processus tumoral pour que leur blocage ait une activité thérapeutique. Il est donc impératif, lorsque l'on passe de la phase pré-clinique aux phases cliniques de l'évaluation des nouvelles thérapeutiques, de pouvoir proposer aux cliniciens de nouveaux outils méthodologiques adaptés à cette nouvelle ère thérapeutique. C'était tout l'enjeu de ce travail de thèse.

REFERENCES

A'Hern, R. P. Widening eligibility to phase II trials: Constant arcsine difference phase II trials. Controlled Clinical Trials. 2004 Jun;25(3), 251-264.

Blackledge G, Lawton F, Redman C, Kelly K. Response of patients in phase II studies of chemotherapy in ovarian cancer: implications for patient treatment and the design of phase II trials. Br J Cancer. 1989 Apr;59(4):650-3.

Bonnefoi H, A'Hern RP, Fisher C, Macfarlane V, Barton D, Blake P, Shepherd JH, Gore ME. Natural history of stage IV epithelial ovarian cancer. J Clin Oncol. 1999 Mar;17(3):767-75.

Brown SR, Gregory WM, Twelves CJ, Buyse M, Collinson F, Parmar M, Seymour MT, Brown JM. Designing phase II trials in cancer: a systematic review and guidance. Br J Cancer. 2011 Jul 12;105(2):194-9. doi: 10. 1038/bjc. 2011. 235. Epub 2011 Jun 28. Review.

Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics. 1995 Dec;51(4):1372-83.

Cadranel J, Beau-Faller M, Mauguen A, Lizard S, Madelaine J, Lansiaux A, Prétet J, Madroszyk A, Chouaid C, Morin F. ASCO Meeting Abstracts May 20 2009: 8079. Biological and clinical prognostic factors in patients with advanced non-small-cell cancer (NSCLC) treated by erlotinib: Preliminary results of the ERMETIC cohort.

Cannistra SA. Phase II trials in journal of clinical oncology. J Clin Oncol. 2009 Jul 1;27(19):3073-6. Epub 2009 May 18. No abstract available.

Cascinu S, Aschele C, Barni S, Debernardis D, Baldo C, Tunesi G, Catalano V, Staccioli MP, Brenna A, Muretto P, Catalano G. Thymidylate synthase protein expression in advanced colon cancer: correlation with the site of metastasis and the clinical response to leucovorin-modulated bolus 5-fluorouracil. Clin Cancer Res. 1999 Aug;5(8):1996-9.

Chang MN, Shuster JJ, Hou W. Improved two-stage tests for stratified phase II cancer clinical trials. Stat Med. 2012 Mar 16. doi: 10. 1002/sim. 5314.

Chen CM, Chi Y. Curtailed two-stage designs with two dependent binary endpoints. Pharm Stat. 2012 Jan-Feb;11(1):57-62. doi: 10. 1002/pst. 496.

Conaway MR, Petroni GR. Bivariate sequential designs for phase II trials. Biometrics. 1995 Jun;51(2):656-64.

Conaway MR, Petroni GR. Designs for phase II trials allowing for a trade-off between response and toxicity. Biometrics. 1996 Dec;52(4):1375-86.

Cressie N, Biele J. A sample-size-optimal Bayesian procedure for sequential pharmaceutical trials. Biometrics. 1994 Sep;50(3):700-11.

de Boo TM, Zielhuis GA. Minimization of sample size when comparing two small probabilities in a non-inferiority safety trial. Stat Med. 2004 Jun 15;23(11):1683-99.

Dingemans AM, Witlox MA, Stallaert RA, van der Valk P, Postmus PE, Giaccone G. Expression of DNA topoisomerase II alpha and topoisomerase IIbeta genes predicts survival and response to chemotherapy in patients with small cell lung cancer. Clin Cancer Res. 1999 Aug;5(8):2048-58.

Elkin EB, Weinstein MC, Winer EP, Kuntz KM, Schnitt SJ, Weeks JC. HER-2 testing and trastuzumab therapy for metastatic breast cancer: a cost-effectiveness analysis. J Clin Oncol. 2004 Mar 1;22(5):854-63.

Estey EH, Thall PF. New designs for phase 2 clinical trials. Blood. 2003 Jul 15;102(2):442-8. Epub 2003 Jan 30.

Fleming TR. One-sample multiple testing procedure for phase II clinical trials. Biometrics. 1982 Mar;38(1):143-51.

Fossella FV, Lee JS, Murphy WK, Lippman SM, Calayag M, Pang A, Chasen M, Shin DM, Glisson B, Benner S, et al. Phase II study of docetaxel for recurrent or metastatic non-small-cell lung cancer. J Clin Oncol. 1994 Jun;12(6):1238-44.

Gazdar AF. Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors. Oncogene. 2009 Aug;28 Suppl 1:S24-31.

Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. J Chronic Dis. 1961 Apr;13:346-53.

Goldman AI, Hannan PJ. Optimal continuous sequential boundaries for monitoring toxicity in clinical trials: a restricted search algorithm. Stat Med. 2001 Jun 15;20(11):1575-89.

Gore ME, Fryatt I, Wiltshaw E, Dawson T. Treatment of relapsed carcinoma of the ovary with cisplatin or carboplatin following initial treatment with these compounds. Gynecol Oncol. 1990 Feb;36(2):207-11.

Guglielmi C, Gomez F, Philip T, Hagenbeek A, Martelli M, Sebban C, Milpied N, Bron D, Cahn JY, Somers R, Sonneveld P, Gisselbrecht C, Van Der Lelie H, Chauvin F. Time to relapse has prognostic value in patients with aggressive lymphoma enrolled onto the Parma trial. J Clin Oncol. 1998 Oct;16(10):3264-9.

Heitjan DF. Bayesian interim analysis of phase II cancer clinical trials. Stat Med. 1997 Aug 30;16(16):1791-802.

Ivanova A, Qaqish BF, Schell MJ. Continuous toxicity monitoring in phase II trials in oncology. Biometrics. 2005 Jun;61(2):540-5.

Jackman DM, Miller VA, Cioffredi LA, Yeap BY, Jänne PA, Riely GJ, Ruiz MG, Giaccone G, Sequist LV, Johnson BE. Impact of epidermal growth factor receptor and KRAS mutations on clinical outcomes in previously untreated non-small cell lung cancer patients: results of an online tumor registry of clinical trials. Clin cancer Res. 2009 Aug 15;15(16):5267-73.

Jones CL, Holmgren E. An adaptive Simon Two-Stage Design for Phase 2 studies of targeted therapies. Contemp Clin Trials. 2007 Sep;28(5):654-61. Epub 2007 Mar 6.

Jung SH, Chang MN, Kang SJ. Phase II cancer clinical trials with heterogeneous patient populations. J Biopharm Stat. 2012;22(2):312-28.

Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, Simes RJ, Chalchal H, Shapiro JD, Robitaille S, Price TJ, Shepherd L, Au HJ, Langer C, Moore MJ, Zalcberg JR. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med. 2008 Oct 23;359(17):1757-65.

Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates?Nat Rev Drug Discov. 2004 Aug;3(8):711-5.

Leblanc M, Rankin C, Crowley J. Multiple Histology Phase II Trials. Clin Cancer Res. 2009 Jul 1;15(13):4256-62.

London WB, Chang MN. One- and two-stage designs for stratified phase II clinical trials. Stat Med. 2005 Sep 15;24(17):2597-611.

Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, Louis DN, Christiani DC,

Settleman J, Haber DA. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 2004 May 20;350(21):2129-39.

Mattson K, Bosquee L, Dabouis G, Le Groumellec A, Pujol JL, Marien S, Stupp R, Douillard JY, Brägas B, Berille J, Olivares R, Le Chevalier T. Phase II study of docetaxel in the treatment of patients with advanced non-small cell lung cancer in routine daily practice. Lung Cancer. 2000 Sep;29(3):205-16.

Morita S, Okamoto I, Kobayashi K, Yamazaki K, Asahina H, Inoue A, Hagiwara K, Sunaga N, Yanagitani N, Hida T, Yoshida K, Hirashima T, Yasumoto K, Sugio K, Mitsudomi T, Fukuoka M, Nukiwa T. Combined survival analysis of prospective clinical trials of gefitinib for non-small cell lung cancer with EGFR mutations. Clin Cancer Res. 2009 Jul 1;15(13):4493-8.

Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004 Jun 4;304(5676):1497-500.

Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, Singh B, Heelan R, Rusch V, Fulton L, Mardis E, Kupfer D, Wilson R, Kris M, Varmus H. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci U S A. 2004 Sep 7;101(36):13306-11.

Phatak AG, Bhatt NM. Estimation of the fraction defective in curtailed sampling plans by attributes. Technometrics 1967; 9:219–228.

Pierga JY, Delaloge S, Espié M, Brain E, Sigal-Zafrani B, Mathieu MC, Bertheau P, Guinebretière JM, Spielmann M, Savignoni A, Marty M. A multicenter randomized phase II study of sequential epirubicin/cyclophosphamide

followed by docetaxel with or without celecoxib or trastuzumab according to HER2 status, as primary chemotherapy for localized invasive breast cancer patients. Breast Cancer Res Treat. 2010 Jul;122(2):429-37.

Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med. 2002 Oct 15;21(19):2917-30.

Pusztai L, Anderson K, Hess KR. Pharmacogenomic predictor discovery in phase II clinical trials for breast cancer. Clin Cancer Res. 2007 Oct 15;13(20):6080-6.

Ray HE, Rai SN. Operating characteristics of a Simon two-stage phase II clinical trial design incorporating continuous toxicity monitoring. Pharm Stat. 2012 Mar-Apr;11(2):170-6.

Roberts TG Jr, Lynch TJ Jr, Chabner BA. The phase III trial in the era of targeted therapy: unraveling the "go or no go" decision. J Clin Oncol. 2003 Oct 1;21(19):3683-95.

Rosell R, Moran T, Queralt C, Porta R, Cardenal F, Camps C, Majem M, Lopez-Vivanco G, Isla D, Provencio M, Insa A, Massuti B, Gonzalez-Larriba JL, Paz-Ares L, Bover I, Garcia-Campelo R, Moreno MA, Catot S, Rolfo C, Reguart N, Palmero R, Sánchez JM, et al. Screening for epidermal growth factor receptor mutations in lung cancer. N Engl J Med. 2009 Sep 3;361(10):958-67.

Roszkowski K, Pluzanska A, Krzakowski M, Smith AP, Saigi E, Aasebo U, Parisi A, Pham Tran N, Olivares R, Berille J. A multicenter, randomized, phase III study of docetaxel plus best supportive care versus best supportive care in chemotherapy-naive patients with metastatic or non-resectable localized non-small cell lung cancer (NSCLC). Lung Cancer. 2000 Mar;27(3):145-57.

Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol. 2005 Mar 20;23(9):2020-7.

Seymour L, Ivy SP, Sargent D, Spriggs D, Baker L, Rubinstein L, Ratain MJ, Le Blanc M, Stewart D, Crowley J, Groshen S, Humphrey JS, West P, Berry D. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. Clin Cancer Res. 2010 Mar 15;16(6):1764-9.

Schaid DJ, Wieand S, Therneau, TM. Optimal two-stage screening designs for survival comparisons. Biometrika 1990; 77:659–663

Shepherd FA, Rodrigues Pereira J, Ciuleanu T, Tan EH, Hirsh V, Thongprasert S, Campos D, Maoleekoonpiroj S, Smylie M, Martins R, van Kooten M, Dediu M, Findlay B, Tu D, Johnston D, Bezjak A, Clark G, Santabárbara P, Seymour L; National Cancer Institute of Canada Clinical Trials Group. Erlotinib in previously treated non-small-cell lung cancer. N Engl J Med. 2005 Jul 14;353(2):123-32.

Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989 Mar;10(1):1-10.

Sleight P. Debate: Subgroup analyses in clinical trials: fun to look at – but don't believe them! Curr Control Trials Cardiovasc Med. 2000;1(1):25-27.

Sposto R, Gaynon PS. An adjustment for patient heterogeneity in the design of two-stage phase II trials. Stat Med. 2009 Sep 10;28(20):2566-79.

Stallard N. Sample size determination for phase II clinical trials based on Bayesian decision theory. Biometrics. 1998 Mar;54(1):279-94.

Stallard N. Approximately optimal designs for phase II clinical studies. J Biopharm Stat. 1998 Jul;8(3):469-87.

Stallard N, Thall PF, Whitehead J. Decision theoretic designs for phase II clinical trials with multiple outcomes. Biometrics. 1999 Sep;55(3):971-7.

Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LH, Benjamin RS. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. Stat Med. 2003 Mar 15;22(5):763-80.

Thall PF, Sung HG. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. Stat Med. 1998 Jul 30;17(14):1563-80.

Thall PF, Simon R, Ellenberg SS. A two-stage design for choosing among several experimental treatments and a control in clinical trials. Biometrics. 1989 Jun;45(2):537-47.

Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. Stat Med. 1995 Feb 28;14(4):357-79.

Thall PF, Cheng SC. Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. Biometrics. 1999 Sep;55(3):746-53.

Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. Biometrics. 1994 Jun;50(2):337-49.

Thall PF, Cheng SC. Optimal two-stage designs for clinical trials based on safety and efficacy. Stat Med. 2001 Apr 15;20(7):1023-32.

Thall PF, Estey EH. A Bayesian strategy for screening cancer treatments prior to phase II clinical evaluation. Stat Med. 1993 Jul 15;12(13):1197-211.

Thall PF, Simon RM, Estey EH. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. J Clin Oncol. 1996 Jan;14(1):296-303.

Thatcher N, Chang A, Parikh P, Rodrigues Pereira J, Ciuleanu T, von Pawel J, Thongprasert S, Tan EH, Pemberton K, Archer V, Carroll K. Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised, placebo-controlled, multicentre study (Iressa Survival Evaluation in Lung Cancer). Lancet. 2005 Oct 29-Nov 4;366(9496):1527-37.

Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. 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. 2000 Feb 2;92(3):205-16.

Thezenas S, Duffour J, Culine S, Kramar A. Five-year change in statistical designs of phase II trials published in leading cancer journals. Eur J Cancer. 2004 May;40(8):1244-9.

Tournoux C, De Rycke Y, Médioni J, Asselain B. Methods of joint evaluation of efficacy and toxicity in phase II clinical trials. Contemp Clin Trials. 2007 Jul;28(4):514-24.

Tournoux-Facon C, De Rycke Y, Tubert-Bitter P. Targeting population entering phase III trials: a new stratified adaptive phase II design. Stat Med. 2011 Apr 15;30(8):801-11.

Tournoux-Facon C, De Rycke Y, Tubert-Bitter P. How a new stratified adaptive phase II design could improve targeting population. Stat Med. 2011 Jun 15;30(13):1555-62.

Van Cutsem E, Lang I, D'haens G et al. ASCO Meeting Abstracts May 20 2008: 2. KRAS status and efficacy in the first-line treatment of patients with metastatic colorectal cancer (mCRC) treated with FOLFIRI with or without cetuximab: The CRYSTAL experience.

Wathen JK, Thall PF, Cook JD, Estey EH. Accounting for patient heterogeneity in phase II clinical trials. Stat Med. 2008 Jul 10;27(15):2802-15.

Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA. 1991 Jul 3;266(1):93-8.

ANNEXES

Available online at www.sciencedirect.com
ScienceDirect
Contemporary Clinical Trials 28 (2007) 514–524
Contemporary Clinical Trials
www.elsevier.com/locate/conclintrial

Methods of joint evaluation of efficacy and toxicity in phase II clinical trials ☆

Caroline Tournoux [a,*], Yann De Rycke [a], Jacques Médioni [b], Bernard Asselain [a]

[a] *Service de Biostatistique, Institut Curie 26 rue d'Ulm, 75248 Paris Cedex 05, France*
[b] *Département de cancérologie médicale, Hôpital Européen Georges-Pompidou, 20,rue Leblanc, 75015 Paris, France*

Received 1 June 2006; accepted 22 January 2007

Abstract

Phase II clinical trials in oncology are usually conducted to evaluate the anti-tumor effect. Because phase I trials are small studies, the maximum tolerated dose of a new drug may not be precisely established and the recommended dose used may lead to excessive toxicity. We investigate the methods proposed by Conaway–Petroni and Bryant–Day allowing early termination of phase II clinical trials and based on joint evaluation of treatment efficacy and safety. Both study designs are computed to minimize the expected accrual under the null hypothesis. As two criteria are considered, the null hypothesis is an area. Each method defines two specific type I error risks. Bryant–Day demonstrate that response and toxicity may be considered as independent ($\Phi=1$). We compare the properties of these two methods with exact calculation according to objective criteria and present one example from a study conducted in France. The two methods differ with regard to the definition of the risks and the assumption of independence. They are similar in terms of expected accruals when $\Phi=1$. Deviations from the assumption of independence induce minor consequences on the type I error risks when the constraint on the type II error risk is less than 15%. Choosing Φ has a minimal impact on expected accrual. Finally, one type I error risk (α_{00}) defined by Conaway–Petroni dramatically increases in the case of deviation from the assumption made on Φ. Due to its robustness in relation to a deviation from the independence assumption, we recommend the use of the Bryant–Day method in clinical practice.

Keywords: Phase II clinical trials; Early stopping rules; Joint efficacy and toxicity; Group sequential design

1. Introduction

Phase II clinical trials in oncology are usually small studies conducted to evaluate the anti-tumor effect of an experimental therapy. For conventional cytotoxic drugs, the primary endpoint is generally clinical response as defined by the Response Evaluation Criteria in Solid Tumors (RECIST) and based on Computed Tomography Scan tumor size measurement [1]. If the new therapy is proven to have sufficient activity, then it may be further studied in a randomized phase III trial. These trials are restricted to patients with specific types of cancer, selected on the basis of the activity of

☆ Supported by a grant from the Fondation pour la Recherche Médicale.
* Corresponding author. Tel.: +33 1 44 32 46 66; fax: +33 1 44 32 40 78.
E-mail address: caroline_tournouxfacon@yahoo.fr (C. Tournoux).

1551-7144/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2007.01.008

the drug in preclinical cancer models, the mechanism of action of the drug, and the activity observed in phase I trials. Several authors have proposed randomization of phase II trials [2,3]. The purpose of randomization in this setting is to minimize the impact of random imbalances in prognostic variables. However, randomized phase II studies are not intended to provide an adequately powered comparison between arms. Therefore, no statistical comparison should be performed.

Ethical concerns that a trial must be stopped early if the experimental treatment appears to be ineffective have led to the development of sequential designs for phase II trials. The first sequential design was proposed by Gehan in 1961 [4] and allows early termination for inefficacy only. Fleming's design [5] and the Triangular Test [6] have been developed to enable early termination of a trial when the treatment is either clearly effective or clearly ineffective. Simon [7] improved Fleming's two-stage design by minimizing either the maximum sample size or the expected sample size under the hypothesis of treatment inefficacy. Ensign developed a three-stage design which integrates Simon's designs [8]. All of these designs are based on a single binary indicator of treatment efficacy while ignoring safety considerations.

Because phase I trials are small studies, the maximum tolerated dose of a new drug may not be clearly established. The recommended dose used in a subsequent phase II trial may lead to excessive toxicity. Moreover, information obtained in phase I may not be directly relevant with respect to patient characteristics or possible interactive effects of multiple agents. Hence, multistage designs that allow for early termination of the study based on joint treatment efficacy and safety have been developed since 1995. Bayesian methods have been proposed by Thall, Simon and Estey [9,10], Thall and Sung [11] and Stallard, Thall and Whitehead [12], while Bryant and Day [13] and Conaway and Petroni [14] have proposed frequentist methods.

In this paper, we investigate and present the two frequentist methods which are an extension of Simon's bivariate endpoint design. We compare their properties using exact calculation and illustrate this work using one example from a study of combined chemotherapy and radiotherapy for head and neck cancer patients conducted in the Institut Curie in Paris, France.

2. Material and methods

2.1. General principles

Response and non-toxicity are binary endpoints. Therapeutic efficacy is evaluated by the parameter P_R and therapeutic safety is evaluated by the parameter P_T, which correspond to the true probability of response and non-toxicity in a given population, respectively. If the true probability of response (or non-toxicity) is less than or equal to a predetermined value P_{R0} (or P_{T0}), the efficacy (or safety) of the treatment will be considered to be insufficient. Otherwise, if both P_R and P_T are greater than P_{R0} and P_{T0}, the treatment will be considered to be sufficiently effective

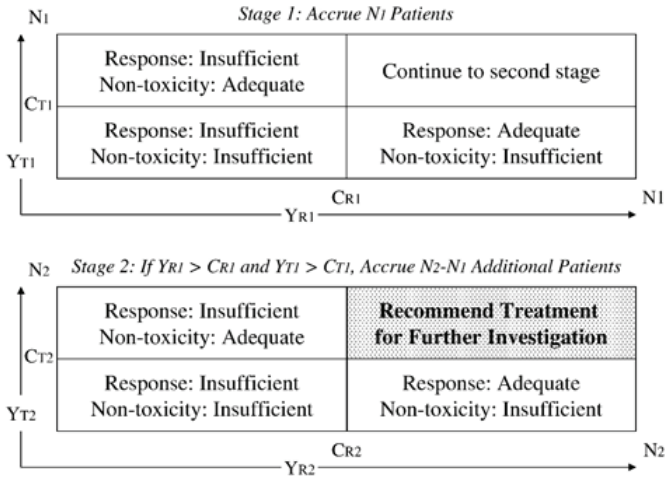

Fig. 1. (Adapted from Bryant & Day). Possible conclusions of a phase II response/toxicity trial. YRk and YTk ($k=1, 2$) are the cumulative number of responses and non-toxicities observed during the first and second stages.

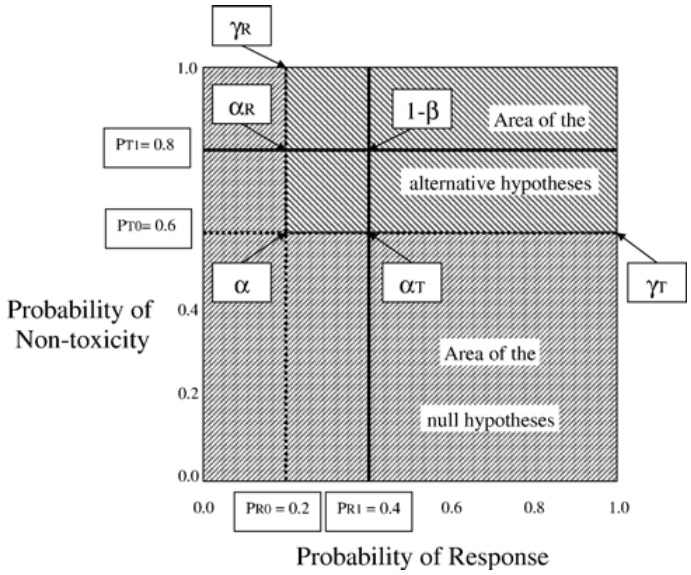

Fig. 2. Graphic representation of the probability to continue treatment evaluation in phase III according to P_R and P_T. Hypotheses for this example are $P_{R0}=0.2$, $P_{R1}=0.4$, $P_{T0}=0.6$ and $P_{T1}=0.8$. The areas of null and alternative hypotheses are defined. Particular probabilities corresponding to type I (α_R, α_T for Bryant & Day and α, γ for Conaway & Petroni) and II (β for both) error points of the two methods are presented.

and safe for further investigation in phase III trials. P_{R1} and P_{T1} are the predetermined rates of response and non-toxicity that we would like to be able to detect if $P_R \geq P_{R1}$ and $P_T \geq P_{T1}$. The predetermined values P_{R0} and P_{T0}.

The methods are built in at least two-stages. We will only consider the case of two-stages and the notation used by Bryant and Day in their article. N_1 patients are included in first stage and Y_{R1} successes and Y_{T1} non-toxicities are observed. If Y_{R1} or Y_{T1} are less than or equal to predetermined values C_{R1} for response and C_{T1} for non-toxicity, the trial is stopped. If both Y_{R1} and Y_{T1} are greater than C_{R1} and C_{T1}, recruitment continues. A total of N_2 patients is included in the trial, among whom Y_{R2} cumulative successes and Y_{T2} cumulative non-toxicities are observed. If Y_{R2} and Y_{T2} are greater than the predetermined values C_{R2} and C_{T2}, investigations continue in phase III trials, otherwise treatment is considered to be ineffective or toxic or both (Fig. 1).

At each stage k ($k=1, 2$), a decision to stop the trial or continue treatment evaluation is taken with associated error risks. Error risks will be presented in Section 2.2 because they are specific to each method.

As we are dealing with two binary variables (response and non-toxicity), we must consider a bivariate binomial distribution with a correlation between the two variables, as the joint distribution depends on the association between response and non-toxicity. The chosen association parameter is the odds-ratio $\Phi = \frac{P_{00}P_{11}}{P_{01}P_{10}}$ (with P_{ij}: $i=0, 1$ for response and $j=0, 1$ for non-toxicity) where P_{00} is the proportion of patients who do not respond and who experience toxicity, P_{01} is the proportion who do not respond and who do not experience toxicity, P_{10} is the proportion who respond and who experience toxicity, and P_{11} is the proportion who respond without toxicity. A positive (or negative) correlation between response and toxicity corresponds to $\Phi<1$ (or $\Phi>1$), since Φ is the odds-ratio associated with response and non-toxicity. The assumption of independence corresponds to $\Phi=1$.

The symbol Φ will be used for the hypothesis made on the correlation structure. Actual correlation structure will be reported by OR for odds-ratio.

2.2. The hypotheses and associated error risks

As two criteria are considered, the null and alternative hypotheses are areas. The authors consider the area of the null hypotheses H_0: $\{(P_R, P_T)|P_R \leq P_{R0}$ or $P_T \leq P_{T0}\}$ and the area of the alternative hypotheses H_1: $\{(P_R, P_T)|P_R > P_{R0}$ and $P_T > P_{T0}\}$. These areas include the particular hypotheses H_{ij}: $P_R=P_{Ri}$ and $P_T=P_{Ti}$ ($i=0, 1$ or $\bullet$ for response and $j=0, 1$ or $\bullet$ for non-toxicity) as defined below.

H_{00}: $P_R=P_{R0}$ and $P_T=P_{T0}$
H_{01}: $P_R=P_{R0}$ and $P_T=P_{T1}$

H_{10}: $P_R = P_{R1}$ and $P_T = P_{T0}$
$H_{0\bullet}$: $P_R = P_{R0}$ and $P_T = 1$
$H_{\bullet 0}$: $P_R = 1$ and $P_T = P_{T0}$
H_{11}: $P_R = P_{R1}$ and $P_T = P_{T1}$

The following risks are associated (Fig. 2):

- α=Upper bound on the probability of recommending a treatment associated with inadequate response and excessive toxicity.
- α_R=Upper bound on the probability of recommending a treatment associated with inadequate response and particular adequate non-toxicity for phase III evaluation.
- α_T=Upper bound on the probability of recommending a treatment associated with particular adequate response and excessive toxicity for phase III evaluation.
- γ=Upper bound on the probability over the limit null hypothesis of recommending a treatment associated with inadequate response or excessive toxicity. This maximum occurs at the point $(P_R, P_T)=(P_{R0}, 1)$ or at $(P_R, P_T)=(1, P_{T0})$. The first point will be noted γ_R and the second point will be noted γ_T.
- β=Upper bound on the probability of not recommending a treatment associated with adequate response and adequate non-toxicity for phase III evaluation.

α, α_R, α_T and γ are type I error risks. β is type II error risk.

Actual (or calculated) risks values will be noted α_{00}, α_{01}, α_{10}, $\gamma_{0\bullet}$, $\gamma_{\bullet 0}$ and $1-\alpha_{11}$ for α, α_R, α_T, γ_R, γ_T and β respectively.

According to B&D method, the optimal design (sample sizes and decision criteria) is that which minimizes the maximal expected sample size under H_{01} or H_{10} ($\max\{E_{01}, E_{10}\}$), controlling the error rates α_R, α_T and β. The authors demonstrate this achievement either uniformly over all possible correlation structures linking response and toxicity, or alternatively, under the assumption of independence between response and toxicity. Designs assuming $\Phi=1$ are available at: (http://biostats.upci.pitt.edu/biostats/ClinicalStudyDesign/Phase2BryantDay.html).

According to C&P the optimal design we chose is that which minimizes the maximal expected sample size under $H_{\bullet 0}$ or $H_{0\bullet}$ ($\max\{E_{\bullet 0}, E_{0\bullet}\}$), satisfying error rates α, γ and β and the assumption made on Φ.

2.3. Comparison of the two methods

We computed exact calculations with S-Plus 6 software (Insightful, Seattle, WA) in order to compare the properties of the two methods. We also applied the methods to the data from one phase II clinical trial conducted at the Institut Curie in Paris.

We first compared study designs obtained with the two methods in order to verify whether they were the same when Φ is assumed to be equal to 1. In order to calculate study designs according to the B&D method, we chose 3 values for P_{R0} (0.05–0.40 and 0.75) and 3 values for P_{T0} (0.05–0.40 and 0.75). Delta (Δ) between P_{R0} (P_{T0}) and P_{R1} (P_{T1}) was fixed equal to 20%. Four errors bounds combinations (α_R–α_T–β) were considered. They were chosen because of their statistical meaning in clinical research: (0.05–0.05–0.05), (0.15–0.15–0.15), (0.05–0.15–0.10), (0.15–0.05–0.10). For each design, we calculated the actual values α_{00} and max ($\gamma_{0\bullet}$, $\gamma_{\bullet 0}$) and used them as α and γ bounds errors to

Table 1
Definition of null and alternative hypotheses according to Bryant & Day and Conaway & Petroni methods

Hypotheses		Errors bounds	Actual error risks	Method
H_{01}	$P_R = P_{R0}$ and $P_T = P_{T1}$	α_R	α_{01}	Bryant & Day
H_{10}	$P_R = P_{R1}$ and $P_T = P_{T0}$	α_T	α_{10}	Bryant & Day
H_{00}	$P_R = P_{R0}$ and $P_T = P_{T0}$	α	α_{00}	Conaway & Petroni
$H_{0\bullet}$	$P_R = P_{R0}$ and $P_T = 1$	γ_R	$\gamma_{0\bullet}$	Conaway & Petroni
$H_{\bullet 0}$	$P_R = 1$ and $P_T = P_{T0}$	γ_T	$\gamma_{\bullet 0}$	Conaway & Petroni
H_{11}	$P_R = P_{R1}$ and $P_T = P_{T1}$	$1-\beta$	α_{11}	Bryant & Day; Conaway & Petroni

Associated bounds errors and actual errors risks are also defined.

determine study designs with the C&P method. We then calculated the differences of expected accruals between the two methods (Table 1).

To evaluate the robustness of the method proposed by B&D, we tested the effect of a deviation from the assumption of independence ($\Phi=1$) on the values of type I and type II error risks. Actual values α_{01}, α_{10} and $1-\alpha_{11}$ were calculated for values of OR varying from 0.01 to 100. We also calculated actual values α_{00} and $\gamma_{0\bullet}$, $\gamma_{\bullet 0}$.

Secondly, we calculated study designs according to the C&P method in order to compare expected accruals under several assumed associations between response and toxicity. P_{R0}, P_{R1} and P_{T0}, P_{T1} values were chosen as previously. Five errors bounds combinations (α–γ–β) were considered: (0.01–0.05–0.05), (0.01–0.05–0.10), (0.01–0.05–0.15), (0.05–0.15–0.05), (0.05–0.15–0.15). We proposed to test five hypotheses on Φ values: 0.01–0.25–1–10–100. For each design, we calculated the expected accrual E_{00}, E_{01}, E_{10} and max ($E_{0\bullet}$, $E_{\bullet 0}$). In order to quantify the impact of Φ on the increase or decrease of expected sample sizes, we defined Θ as the ratio between expected accrual when $\Phi\neq 1$ and the expected accrual when $\Phi=1$.

$$\Theta=\frac{E_{ij}(\Phi\neq 1)}{E_{ij}(\Phi=1)}\quad (i=0,\ \text{or}\ \bullet\ \text{and}\ j=0,\ 1\ \text{or}\ \bullet).$$

We also tested the robustness of the C&P method in the case of a deviation from an assumed value of Φ. We calculated actual values α_{00}, $\gamma_{0\bullet}$, $\gamma_{\bullet 0}$ and $1-\alpha_{11}$ for values of OR varying from 0.01 to 100 when assumed values of Φ were equal to 0.25 or 10. We also calculated actual values α_{01} and α_{10}.

3. Results

3.1. Difference of the expected accruals (E_{ij}) under the assumption of independence ($\Phi=1$) between the two methods and effect of a deviation from the assumption of independence on type I and type II error risks

The absolute average difference of expected accruals E_{00}, E_{10} and E_{01} is less than 2 patients for the various study designs studied (1.257, 1.039 and 1.044 respectively). This difference is on average of 3.093 under H_{11}. Differences between the two methods are therefore minimal. These results were expected, as we chose risks as compatible as possible between the two methods. The slight difference is related to selection of the optimal design, specific to each method, as indicated above.

To evaluate the robustness of the method proposed by B&D, we then tested the effect of a deviation from the assumption of independence on type I and type II error risks, for study designs calculated according to the B&D method. When error bound $\beta<15\%$, constraints are respected when OR $\neq 1$. α_{01} and α_{10} decrease when OR is close to 0 and increase when OR is close to 100. α_{11} is stable, increasing slightly as OR increases, particularly for designs with $P_{R0}=P_{T0}$. When $\beta=15\%$, constraints may not be respected for hypotheses proposed in clinical practice. An example is

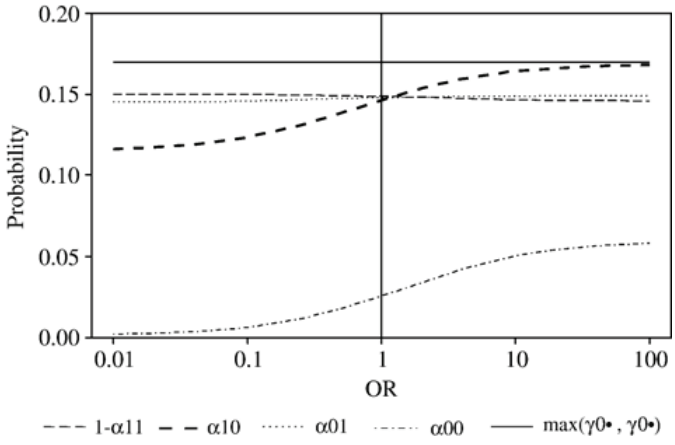

Fig. 3. Actual risks values (Probability) α_{00}, α_{01}, α_{10}, $\gamma_{0\bullet}$, $\gamma_{\bullet 0}$ and $1-\alpha_{11}$ in the case of a deviation from the assumption of independence ($\Phi=1$) for the following Bryant & Day design: $P_{R0}=0.40$ and $P_{T0}=0.75$, $\Delta=20\%$ and risks α_R, α_T and β equal to 15%.

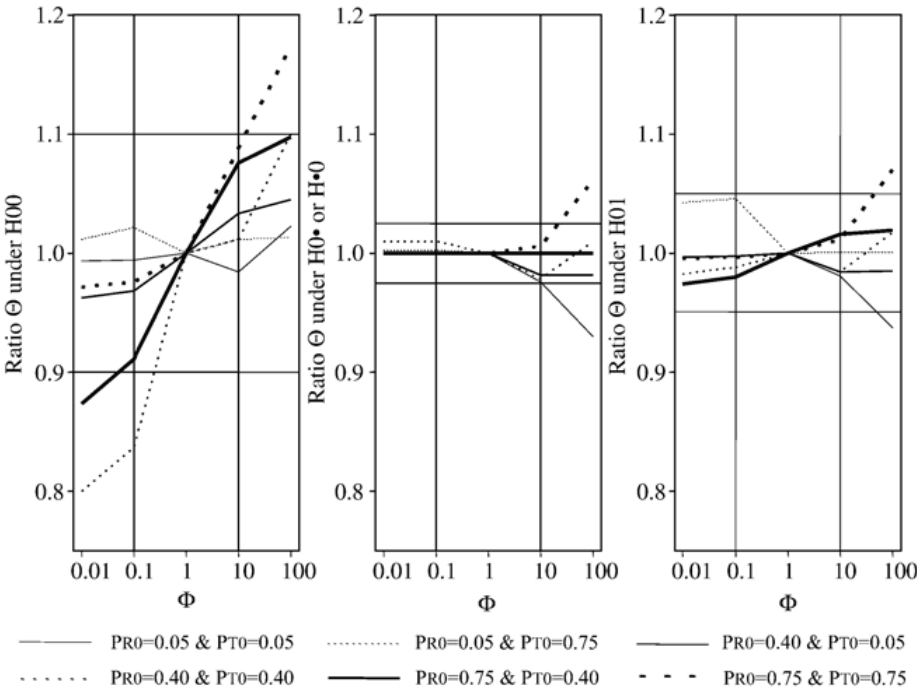

Fig. 4. Ratio between expected accrual when $\Phi \neq 1$ and expected accrual when $\Phi = 1$ for Conaway and Petroni designs under H_{00}, under $H_{0\bullet}$ or $H_{\bullet 0}$ and H_{01} according to several prior values of Φ (0.01, 0.25, 1, 10, 100), for $\alpha = 0.01$, $\gamma = 0.05$ and $\beta = 0.10$ and for six situations of rates.

given Fig. 3 which presents the calculated risks values of α_R, α_T and β (Probability), noted α_{01}, α_{10} and $1-\alpha_{11}$ respectively, according to actual correlation structure between response and toxicity (OR). Values of α_{00}, $\gamma_{0\bullet}$ and $\gamma_{\bullet 0}$ for α_{00}, $\gamma_{0\bullet}$ and $\gamma_{\bullet 0}$ respectively have also been calculated for information. For $P_{R0} = 0.40$–$P_{T0} = 0.75$, α_{10} increases up to 0.168 as OR increases. B&D demonstrate that true type I error risks could be inflated to as much as $\alpha_R/(1-\beta)$ and $\alpha_T/(1-\beta)$.

3.2. Expected accruals (E_{ij}) under an assumed association between response and toxicity and effect of a misspecification of Φ on type I and type II error risks

To test the impact of Φ on the increase or decrease of expected sample sizes for study designs calculated according to the C&P method, we calculated Θ for each design of risks and rates. For moderate assumed values of Φ between 0.1 and 10, the ratio Θ varies between 0.9 and 1.1 (less than 10% of variation), except when $P_{R0} = 0.40$ and $P_{T0} = 0.40$ (Fig. 4). In this situation, only 32.11 patients are expected when $\Phi = 0.1$ versus 38.50 when $\Phi = 1$ (more than 15% of variation). Expected sample sizes $E_{\bullet 0}$ or $E_{0\bullet}$ and E_{01} are also similar to the situation where $\Phi = 1$ (variation less than 2.5 and 5%, respectively) for Φ between 0.1 and 10 (E_{10} is similar to E_{01} and is then not represented). These results support the B&D point of view by minimizing the influence of Φ on expected accrual.

α_{00} increases dramatically in the case of a deviation from the assumption made on Φ, particularly when $P_{R0} = P_{T0}$ and even for β less than 15%. α_{00} increases when actual OR increases. For example, with a prior value of $\Phi = 0.25$, α_{00} will be 6, 11 and 25 times higher if OR is actually equal to 4, 10 and 100, respectively (Fig. 5a).

The power may also not be maintained under the constraint. For a prior value of $\Phi = 0.10$, $1-\alpha_{11}$ increases by more than 17.5% when OR is less than 10 (Fig. 5b).

On the contrary, the risk corresponding to max ($\gamma_{0\bullet}$, $\gamma_{\bullet 0}$) is stable regardless of the prior hypothesis made on Φ and even in the case of a large deviation from the hypothesis (Fig. 5a and b).

3.3. Carbora trial

In order to illustrate the results, an application of the methods on real data from a clinical trial is proposed. We present conclusions from the classical sequential method used in practice and from Simon's Optimal method because

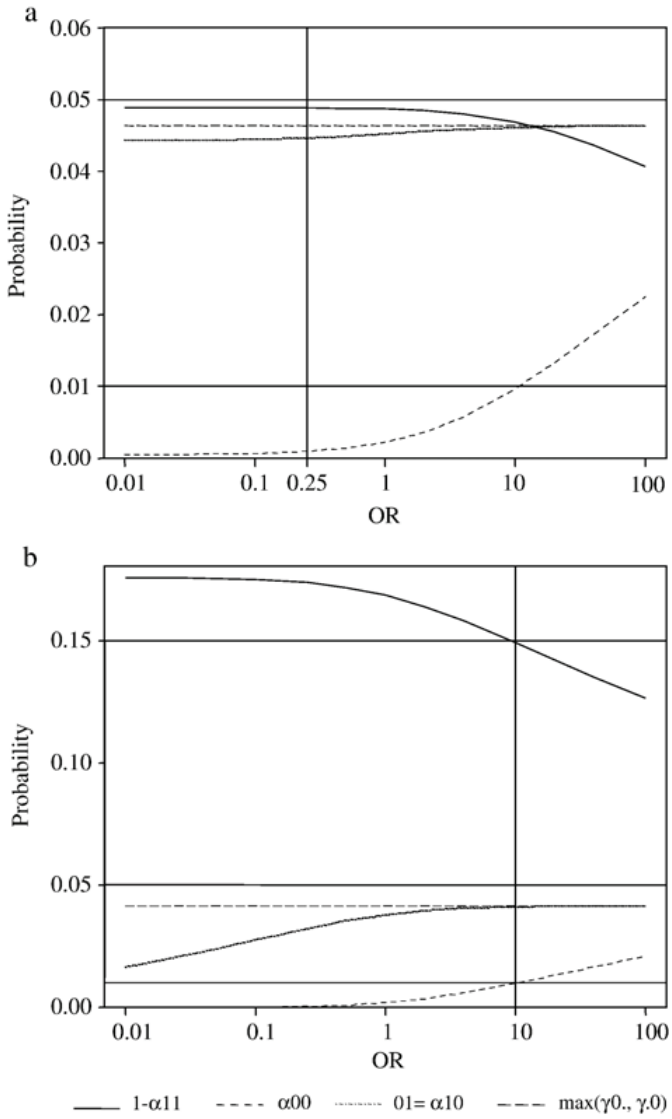

Fig. 5. Actual risks values (Probability) α_{00}, α_{01}, α_{10}, $\gamma_{0\bullet}$, $\gamma_{.0}$ and $1-\alpha_{11}$ in the case of a deviation from the assumption made on Φ ($\Phi=0.25$, a and $\Phi=10$, b) according to different hypotheses on risks (α, γ and β) and rates for Conaway and Petroni designs. a: α=1%; γ=5%; β=5%; P_{RO}=0.75; P_{TO}=2.75. b: α=1%; γ=5%; β=15%; P_{RO}=0.40; P_{TO}=0.40.

B&D and C&P are an extension of this method. We consider α_R equal to classical type I error risk and arbitrarily to α_T. The risks α and γ as defined by C&P are calculated with the same methodology as in 2.3. The power $1-\beta$ is that used in the classical method.

Table 2
Carbora trial

Methods P_{R0}–P_{R1}	Stage	Theoretical cumulative accrual	Bounds				$E(n\|H_0)$	α β	Conclusion	Actual included accrual
				$\leq a$		$\geq b$				
Fleming 0.55–0.75	1	15		7	*9*	15	26.62	0.05	Inefficacy	30
	2	30	**18**	19		23		0.13		
	3	45		30		31				
Simon 0.55–0.75	1	17		9	*10*		28.18	0.05	Impossible†	44
	2	44	18†	29		30		0.13		

Designs according to classical methods. We present theoretical cumulative accrual at each stage, inferior (a) and superior (b) bounds and cumulative numbers of observed responses (in bold, inclusion is stopped; in italic, inclusion continues), expected accrual under H_0, type I and type II error risks (α and β), conclusion and actual number of patients included at the end of the trial (†: only 18/32 responses were observed).

Table 3
Carbora trial

Methods P_{R0}–P_{R1}–P_{T0}–P_{T1}	Stage	Theoretical cumulative accrual	Bounds $\leq C_{Ri}$ $\leq C_{Tj}$		$E(n\|H_{00})$	α α_R α_T γ	β	Conclusion	Actual included accrual
Bryant–Day	1	19	**10**	11	22.62	–	0.13	Inefficacy	19
0.55–0.75–0.82–0.97			**14**	16		0.05		Toxicity	
	2	56		36		0.05			
				50		–			
Conaway–Petroni (ϕ=1/9)	1	19	**10**	11	20.57	0.002	0.13	Inefficacy	19
0.55–0.75–0.82–0.97			**14**	16		–		Toxicity	
	2	45		28		–			
				40		0.11			
Conaway–Petroni (ϕ=1)	1	19	**10**	11	22.62	0.002	0.13	Inefficacy	19
0.55–0.75–0.82–0.97			**14**	16		–		Toxicity	
	2	56		36		–			
				50		0.05			
Conaway–Petroni (ϕ=9)	1	19	**10**	11	25.47	0.002	0.13	Inefficacy	19
0.55–0.75–0.82–0.97			**14**	16		–		Toxicity	
	2	63		39		–			
				58		0.08			

Designs according to B & D and C & P methods. We present theoretical cumulative accrual at each stage, bounds for response and for toxicity (C_{Ri} and C_{Tj} with i, j=1, 2) and cumulative numbers of observed responses and non-toxicities (in bold, inclusion is stopped), expected accrual under H_{00}, type I and type II error risks (α, α_R, α_T, γ and β according to the corresponding method), conclusion and actual number of patients included at the end of the trial.

"Carbora" was a phase II clinical trial conducted at the Institut Curie. The main objective was to test whether a concomitant combination of conventional radiotherapy and low-dose Carboplatin could achieve a response rate greater than 55% (P_{R0}) in head and neck cancer. A response rate equal to 75% was considered to be a clinically meaningful improvement (P_{R1}). Given a type I error of 5% and a power of 87%, a Fleming's 3-step procedure was used with sequential analysis every 15 patients.

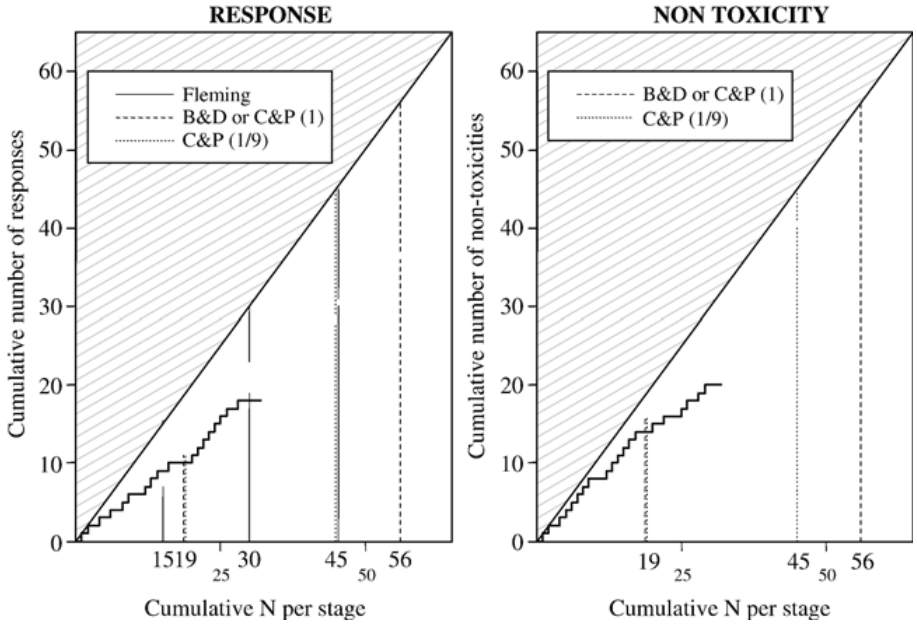

Fig. 6. "Carbora" trial: Cumulative number of observed responses and non-toxicities according to the cumulative number of patients (N) included at each stage. We present stopping bounds for the Fleming, Bryant–Day (B&D) and Conaway–Petroni (C&P) methods (Φ=1, Φ=1/9). When the path of responses (or non-toxicities) crosses the bounds, the trial is stopped for inefficacy or toxicity.

The trial started in 1995. In 1997, at the 2nd analysis involving 30 patients, only 18 responses were observed. 20 responses would have been necessary to continue in step 3. Patient inclusion was therefore stopped and the treatment was considered to be ineffective. According to the Optimal Simon method, 44 patients would have been included to reach a conclusion. As we only had data on 32 patients, no conclusion could be reached with this method (Table 2).

Results according to the B&D and C&P methods are presented in Table 3. P_{T0} is the minimal acceptable non-toxicity rate and P_{T1} is the rate we would like to be able to detect if $P_T \geq P_{T1}$. For this study, non-toxicity rate P_{T0} was based on the definition of the Maximum Tolerated Dose (MTD) in phase I trials. It was considered that a dose producing severe toxicity in 18% of patients was acceptable for further development. P_{T1} was based on results of the literature [15,16]. P_{T0} and P_{T1} were stated equal to 0.82 and 0.97, respectively. Deaths related to treatment and grade 3 or 4 toxicities are considered to be toxic events. To make a hypothesis on Φ, we estimated the odds-ratio observed (OR) in the trial. In this trial, OR = 9, which means that responses are more likely to be associated with non-toxicities. In fact, during the trial observed toxicities have led to discontinuation of treatment. As response criteria were not available for these drop-outs, they have been considered in practice to be failures for this criterion, which has probably completely changed the value of OR. Indeed, as explained in 2.1, a positive correlation between response and toxicity corresponds to $\Phi < 1$. In oncology trials, such a correlation is probably more often expected. We also propose to test the value $\Phi = 1/\mathrm{OR}$.

The two methods result in the same designs when $\Phi = 1$ (Table 3).

The two methods can reach a conclusion for both inefficacy and toxicity at the end of the first stage, after inclusion of 19 patients, saving 11 and 25 patients compared to the number of patients that would have been included with the Fleming and Simon methods, respectively. Cumulative responses and non-toxicities are presented in Fig. 6.

If the hypothesis made on Φ had been 1/9, constraints would have been $\alpha = 0.002$, $\gamma = 0.11$ and $\beta = 0.13$. As in fact OR = 9, α_{00} is close to 1.64%.

4. Discussion

Some drugs tested in phase II, either alone or in combination, present a possibility of high toxicity, because the recommended dose at the end of phase I may not have been correctly estimated as only a few patients would have been included. Response and toxicity should therefore both be taken into account when planning phase II clinical trials. B&D and C&P proposed this type of method 10 years ago.

After comparison, the B&D method appears to differ from the C&P method according to five points:

(i) Type I error risks and hypothesis on parameter Φ are specific to each method. The choice is also firstly conceptual, secondly based on the clinicians' understanding of type I error risk and thirdly on the possibility to determine prior Φ.
(ii) Expected accruals are similar between the two methods in the case of independence ($\Phi = 1$).
(iii) Deviations from the assumption of independence have little impact on B&D type I and type II risks values, only when $\beta = 15\%$.
(iv) Specifying a priori values of Φ has a minimal impact on expected accruals.
(v) Large variations in α_{00} and $1 - \alpha_{11}$ values are observed in the case of a deviation from the assumption made on Φ. However, the aim in phase II trials is to minimize the beta risk (i.e. minimize the false-negative rate) in order to ensure that a useful drug will be studied in phase III. The beta risk increases when OR < 1 and this situation is probably the rule with cytotoxic drugs.

Due to its robustness in relation to a deviation from the independence assumption, we therefore recommend the Bryant and Day method in clinical practice.

Expected numbers of patients to be included with these bivariate methods are slightly higher than those proposed by the Simon or Fleming methods. An overview of such a comparison has been described by B&D [13]. As explained by the authors, "the cost of jointly considering both response and toxicity can be considerable. On the other hand, it can be argued that these same costs apply even if toxicity considerations were to be taken into account in an informal or post hoc manner. Thus, if a study was designed with sample sizes based only on clinical response, the true Type II error rate may be inflated over its nominal level by virtue of the fact that the trial could be terminated due to apparent excessive toxicity. Conversely, at the conclusion of the trial, less may be known about the treatment's toxicity than is required prior to recommending it for Phase II testing."

These methods are not widely used at the present time. We reviewed 399 phase II trials published between January 2003 and April 2005 in *Journal of Clinical Oncology*, *Cancer*, *Annals of Oncology* and *British Journal of Cancer*. We performed a Medline search using the following "M.E.S.H" words: "phase II", "clinical trials". All trials in oncology on adults or children were reviewed. On 399 articles, only one trial published in *Cancer* in December 2003 was conducted with the B&D method [17].

However, these methods present a considerable ethical value in oncology, especially in four situations:

Elderly populations who are often excluded from trials because of their poor performance status and previous treatment.

Children for whom treatment tolerance is often better than in adults, but which must be more precisely studied. Moreover, in these trials, population may be heterogeneous because of age.

Orphan diseases because of recruitment difficulties that would be simpler if only one trial was conducted with both criteria.

Biotherapies for which classical phase I trials are probably inappropriate because of the lack of a linear relationship between dose and toxicity.

However, several criticisms concerning these methods can be formulated. First, if toxicity leads to discontinue the treatment, response may be obtained with difficulties. If toxic event is the death, then response will completely be censored. It will modify the correlation structure between the two criteria. This was observed in the Carbora trial. In such circumstances, such methods based on joint evaluation of treatment efficacy and toxicity may not be appropriate. Secondly, analysis of toxic events is a grouped sequential analysis, at the same time as response. Continuous analysis would probably be more appropriate to severe adverse events. Ivanova [18] proposes recommendations on how to construct stopping rules using Pocock stopping boundaries [19]. Adaptations of the Wald's Sequential Probability Ratio Test [20] or Risk Spending Function of Lan and DeMets [21] are also possible. However, at this time, all these methods lead to independent decisions, one concerning response and one concerning toxicity, making global control of error risks difficult. Finally, toxicity relies on investigator assessment and is then probably subject to variability. The investigators use validated toxicity scales such as the World Health Organization standard toxicity scale. Furthermore, they should also make clear in the protocol the expected adverse events. A convenient insight could be to distinguish between treatment and disease progression related toxicities [22].

5. Conclusion

The methods developed by Bryant & Day and Conaway & Petroni are not widely used at the present time, possibly because of their relatively recent publication and the lack of algorithm or tables for the Conaway & Petroni method in contrast with the Bryant & Day for which an algorithm is available at the address given below.

Because of the difficulty to determine Φ and due to the robustness in relation to a deviation from the independence assumption, we recommend the Bryant and Day method in clinical practice, as the Conaway and Petroni method may fail to control risks under the constraints in the case of a deviation from the assumption of the value made on Φ.

Finally, these methods were initially developed in oncology, but are also applicable to other medical fields and at different stages of development of a new treatment, particularly in phase III [23].

References

[1] 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 2000;92:205–16.
[2] Buyse M. Randomized designs for early trials of new cancer treatments — an overview. Drug Inf J 2000;34:387–96.
[3] Van Glabbeke M, Steward W, Armand JP. Non-randomised phase II trials of drug combinations: often meaningless, sometimes misleading. Are there alternative strategies? Eur J Cancer 2002;38:635–8.
[4] Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. J Chronic Dis 1960;13:346–53.
[5] Fleming TR. One-sample multiple testing procedure for phase II clinical trials. Biometrics 1982;38:143–51.
[6] Bellissant E, Benichou J, Chastang C. Application of the triangular test to phase II cancer clinical trials. Stat Med 1990;9:907–17.
[7] Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials 1989;10:1–10.
[8] Ensign LG, Gehan EA, Kamen DS, Thall PF. An optimal three-stage design for phase II clinical trials. Stat Med 1994;13:1727–36.
[9] Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. Stat Med 1995;14:357–79.

[10] Thall PF, Simon RM, Estey EH. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. J Clin Oncol 1996;14:296–303.
[11] Thall PF, Sung HG. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. Stat Med 1998;17:1563–80.
[12] Stallard N, Thall PF, Whitehead J. Decision theoretic designs for phase II clinical trials with multiple outcomes. Biometrics 1999;55:971–7.
[13] Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics 1995;51:1372–83.
[14] Conaway MR, Petroni GR. Bivariate sequential designs for phase II trials. Biometrics 1995;51:656–64.
[15] Dieras V, Extra JM, Bellissant E, et al. Efficacy and tolerance of vinorelbine and fluorouracil combination as first-line chemotherapy of advanced breast cancer: results of a phase II study using a sequential group method. J Clin Oncol 1996;14:3097–104.
[16] Zambelli A, Robustelli della Cuna FS, Ponchio L, et al. Four-day infusion of fluorouracil plus vinorelbine as salvage treatment of heavily pretreated metastatic breast cancer. Breast Cancer Res Treat 2000;61:241–7.
[17] Garaventa A, Luksch R, Biasotti S, et al. A phase II study of topotecan with vincristine and doxorubicin in children with recurrent/refractory neuroblastoma. Cancer 2003;98:2488–94.
[18] Ivanova A, Qaqish BF, Schell MJ. Continuous toxicity monitoring in phase II trials in oncology. Biometrics 2005;61:540–5.
[19] Pocock S. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191–9.
[20] Wald A. Sequential analysis. New York: Wiley; 1947.
[21] Lan K, DeMets D. Discrete sequential boundaries for clinical trials. Biometrics 1983;70:659–63.
[22] Brundage MD, Pater JL, Zee B. Assessing the reliability of two toxicity scales: implications for interpreting toxicity data. J Natl Cancer Inst 1993;85:1138–48.
[23] Letierce A, Tubert-Bitter P, Kramar A, Maccario J. Two-treatment comparison based on joint toxicity and efficacy ordered alternatives in cancer trials. Stat Med 2003;22:859–68.

Research Article

Statistics in Medicine

Received 18 August 2009, Accepted 4 November 2010 Published online 12 January 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4148

Targeting population entering phase III trials: A new stratified adaptive phase II design

Caroline Tournoux-Facon,[a,b,c,d] Yann De Rycke[e*†] and Pascale Tubert-Bitter[c,d]

The primary goal of phase II studies is to assess the efficacy of the new treatment in order to decide whether it has sufficient activity to warrant further evaluation in a phase III comparative trial. However, many adequately conducted phase II trials are negative leading to termination of drug development. Heterogeneity of the population is often considered to be a cause of treatment effect dilution. One approach to determine the sensitive subpopulation is to conduct several phase II trials, one in each specific subset of patients. This option might unethically increase the number of non-sensitive patients under evaluation. Adaptive two-stage designs have been recently proposed. London and Chang proposed a global one-sample test for response rates for stratified phase II clinical trials, whereas Jones and Holmgren proposed an adaptive design that allows preliminary determination of efficacy that may be restricted to a specific subpopulation defined by biomarker status. These two methods do not allow early termination for efficacy in one or several subgroups as they are extensions of the Simon design. The authors propose an alternative method to deal with stratification in phase II clinical trials and identification of the best target population. This method is based on the multiple-stage Fleming design allowing for early stopping rules for either efficacy or inefficacy. It also integrates a procedure testing whether treatment effects are similar or heterogeneous between the two groups. The operating characteristics of this method were compared with those of a standard Fleming design using exact binomial probabilities. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: adaptive phase II design; target population; stratification

1. Introduction

In oncology, the efficacy of a new treatment during a phase II trial is often evaluated using a binary outcome (response or failure). Minimal acceptable probability of response is set up to determine whether it is worthwhile to perform further investigations with this drug. In a multistage design, observed response rate is computed on accumulating data from patients included sequentially at pre-determined stages. If this observed response rate is sufficiently large (or small), the study is terminated and the decision to move on to phase III testing (or to eliminate this treatment from further consideration) is then made. Multi-stage designs are popular because they have been shown to result in the correct decision with the same accuracy as a one-stage design but with smaller average sample sizes. Although more stages in a multi-stage design are associated with better performance in terms of smaller average sample size, the greatest incremental gains are achieved with fewer stages, i.e. the greatest gain in average sample size is observed when moving from a one-stage design to a two-stage design with a progressively decreasing gain for subsequent stages [1]. This assessment and the logistic difficulties of

[a]*INSERM CIC P-0802, University Hospital of Poitiers, France*
[b]*Epidemiology and Biostatistics, University of Poitiers, France*
[c]*Inserm, CESP Centre for research in Epidemiology and Population Health, U1018, F-94807 Villejuif, France*
[d]*Univ Paris-Sud, UMRS 1018, F-94807 Villejuif, France*
[e]*Service de Biostatistique, Institut Curie, Paris, France*
[*]*Correspondence to: Yann De Rycke, Service de Biostatistique, Institut Curie, 26 rue d'Ulm, 75005 Paris, France.*
[†]*E-mail: yann.de-rycke@curie.net*

setting up multi-stage designs have made two-stage designs especially attractive for phase II clinical trials.

Unfortunately, many well-conducted phase II clinical trials are negative and development of promising treatments is erroneously stopped. Heterogeneity is often claimed as a cause of treatment effect dilution: non-stratified two-stage designs are conducted in pooled clinically or biologically heterogeneous patient cohorts. To confirm greater benefit in an identified subpopulation, one of the investigators' options is to perform independent studies with its own design for each subpopulation, resulting in an increased total number of patients under evaluation, while it is ethically critical to minimize the number of patients who do not benefit from the new drug. One of the challenges of future new drug investigation is the development of new trial designs able to identify a potential sensitive subpopulation, not after but during the study. Stratified adaptive Simon two-stage designs have recently been proposed. London and Chang developed a one-sample test for response rates for stratified phase II clinical trials [2]. This global test is based on the difference of the observed total number of responses across all strata and the corresponding expected number of responses under the null hypothesis. Jones and Holmgren proposed a design that allows for determination of efficacy that may be restricted to a specific subpopulation defined by biomarker status [3]. However, these two methods do not allow early termination in stage 1 for efficacy in one subgroup.

In this paper, we propose a new stratified adaptive phase II design based on the well-known two-stage Fleming design [4] to determine whether a new drug provides benefit to at least one subpopulation of patients. The objective is to stop drug development in the whole population at the first or second stage and to continue with only one subpopulation in the second stage or phase III if only one subpopulation benefits from the treatment. For this purpose, we have developed a new adaptive method in which the second-stage sample size and decision rules depend on the observed response rates at the first stage. The structure of the paper is as follows: we formulate the problem in Section 2, and present the results in Section 3, with a discussion in Section 4.

2. Material and methods

2.1. Assumptions and notations

We will assume that two subpopulations $i=1$ or 2 are predefined and that patients enter the clinical trial sequentially during stage $s=1$ or 2. The sample sizes for the first and second stages are denoted by n_{i1} and n_{i2} in each subpopulation i. We assume that the ratio between the two subpopulations may differ from 1 but is constant between the two stages and *a priori* defined as $n_{2s}=w\times n_{1s}$ (1).

The binary outcome for the jth patient in subpopulation i and stage s is denoted by X_{isj}, where $X_{isj}=1$ in the case of response and 0 in the case of failure. The cumulative number of responses within a given population at a given stage is denoted by $R_{is}=\sum_{j=1}^{nis} X_{isj}$.

For the problem tested, the probability of response in subpopulation i below which the investigational medicinal product is declared to be a low-activity drug is denoted by π_{0i}.

The null hypothesis is denoted by H_0: $\pi_1\leqslant\pi_{01}$ and $\pi_2\leqslant\pi_{02}$ and the alternative hypothesis is denoted by H_1: $\pi_1>\pi_{01}$ or $\pi_2>\pi_{02}$ (Figure 1).

π_{1i} is defined as the lowest probability of success that we would like to detect for subpopulation i and $\Delta_i=\pi_{1i}-\pi_{0i}$. π_{01} may be different from π_{02} and Δ_1 may be different from Δ_2 as well.

2.2. General principles of the stratified adaptive Fleming design

2.2.1. Determination of design parameters. Let π_0 be the marginal probability of response under the null hypothesis $H_{00:}\pi_1=\pi_{01}$ and $\pi_2=\pi_{02}$.

$$\pi_0=P(X=1|i=1)\times P(i=1)+P(X=1|i=2)\times P(i=2).$$

Using formula (1), we deduce

$$\pi_0=\frac{\pi_{01}\times(n_{11}+n_{12})+w\times(n_{11}+n_{12})\times\pi_{02}}{(n_{11}+n_{12})+w\times(n_{11}+n_{12})}=\frac{\pi_{01}+w\times\pi_{02}}{1+w}.$$

Likewise, we define $\Delta=(\Delta_1+w\times\Delta_2)/(1+w)$ and $H_{11}:\pi_1=\pi_{11}$ and $\pi_2=\pi_{12}$.

For π_0, Δ and given prespecified type I and type II errors, α and β, the cumulative number of patients at each stage (n_1,n_2) and the thresholds used for decision rules (a_1,b_1,b_2), are determined by

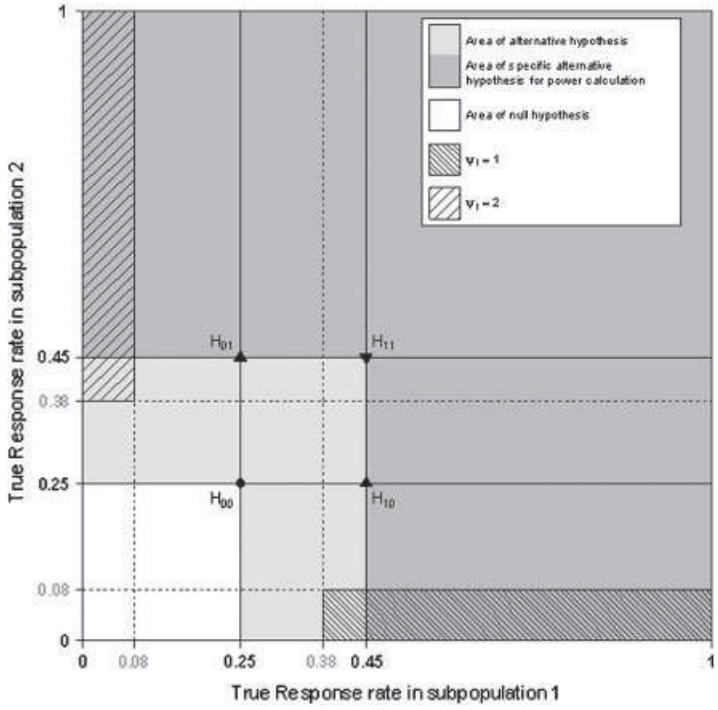

Figure 1. Graphic representation of the areas of null and alternative hypotheses, of the particular hypotheses H_{00}, H_{01}, H_{10} and H_{11} and of the areas of detection of $\Psi_1=1$ or $\Psi_1=2$. Hypotheses for this example are $\pi_{0i}=0.25$, $\Delta=0.2$, $\gamma=0.6$. Lower and upper boundaries of intervals of probability for the first stage of the design (IP_{i1}) are in light grey.

iterations using a Fleming two-stage procedure. According to the classical Fleming design, a_1, b_1, b_2 are boundaries leading to the following decision: At the first stage, if cumulated number of response is smaller than a_1 or greater than b_1 then the trial is stopped for inefficacy or efficacy. Otherwise, inclusion continues to stage 2. The sample size in each subpopulation i and at each stage s $(n_{11}, n_{12}, n_{21}, n_{22})$ are deduced from (1).

2.2.2. Identification of heterogeneity of responses between the two subpopulations at stage 1. Let IP_{is} be a symmetric interval of probability around π_{0i} at each stage s with length at least equal to $(1-\gamma)$. Only approximate symmetry is observed due to binomial calculation.

The determination of heterogeneity in the responses of trial subpopulations at stage s is denoted by Ψ_s.

At the end of the first stage, if R_{11} is less than (greater than) the lower (upper) boundary of IP_{11} and R_{21} is greater than (less than) the upper (lower) boundary of IP_{21}, then the responses between the two subpopulations can be considered to be heterogeneous (Figure 1). In all other situations, the treatment effect is considered to be similar in the two subpopulations and no subpopulation should be selected. At the end of stage 1, three situations may occur:

$\Psi_1=0$: No identification of heterogeneity of responses between the two subpopulations.
$\Psi_1=1$: Identification of heterogeneity of responses in favour of subpopulation 1 (drug development will continue in subpopulation 1 only).
$\Psi_1=2$: Identification of heterogeneity of responses in favour of subpopulation 2 (drug development will continue in subpopulation 2 only).

As IP_{is} intervals are calculated under H_{00}, the combined probability to continue drug evaluation with one subpopulation when the drug is ineffective in both subpopulations is given by:

$$P\left(\frac{R_{11}}{n_{11}} \geqslant \text{upper boundary } IP_{11}\right) \times P\left(\frac{R_{21}}{n_{21}} \leqslant \text{lower boundary } IP_{21}\right)$$
$$+P\left(\frac{R_{11}}{n_{11}} \leqslant \text{lower boundary } IP_{11}\right) \times P\left(\frac{R_{21}}{n_{21}} \geqslant \text{upper boundary } IP_{21}\right) = \frac{\gamma^2}{2}.$$

Table I. Decision rules at the end of first stage, after inclusion of $n_{11}+n_{21}$ patients.

	$R_{11}+R_{21}\leqslant a_1$	$R_{11}+R_{21}\in]a_1-b_1[$	$R_{11}+R_{21}\geqslant b_1$
$\Psi_1=1$	C_1I_2	C_1I_2	E_1I_2
$\Psi_1=0$	I_1I_2	C_1C_2	E_1E_2
$\Psi_1=2$	I_1C_2	I_1C_2	I_1E_2

I_i = inefficacy of subpopulation i; E_i = efficacy of subpopulation i; C_i = continue to stage 2 with subpopulation i.

Table II. Cumulative number of response and patients included in the trial and in the final decision rule according to the decision at stage 1.

	Cumulative number of patients included in the trial	Cumulative number of responses (R_d)/patients included in the final decision rule
Stop stage 1	$n_{11}+n_{21}$	$R_{11}+R_{21}/n_{11}+n_{21}$
Continue to stage 2 with		
Subpopulation 1	$n_{11}+n_{21}+n_{2F1}$	$R_{11}+R_{2F1}/n_{11}+n_{2F1}$
Entire population	$n_{11}+n_{21}+n_{12}+n_{22}$	$R_{11}+R_{21}+R_{12}+R_{22}/n_{11}+n_{21}+n_{12}+n_{22}$
Subpopulation 2	$n_{11}+n_{21}+n_{2F2}$	$R_{21}+R_{2F2}/n_{21}+n_{2F2}$

2.2.3. Conclusions at the end of stage 1. The nine situations are presented in Table I. They depend on the combined information from Ψ_1 and the cumulative number of stage 1 responses $R_{11}+R_{21}$ and raise the following comments: if $\Psi_1=1$ and $R_{11}+R_{21}\leqslant a_1$, drug evaluation will continue to stage 2 for subpopulation 1 and will be stopped in stage 1 for subpopulation 2. In a classical Fleming design, drug evaluation would be stopped for the entire population. At this stage, there is some evidence that subpopulation 1 may be sensitive and that the global inefficacy may be exclusively due to subpopulation 2, which clearly does not benefit. If $\Psi_1=1$ and $R_{11}+R_{21}\geqslant b_1$, drug evaluation will continue to phase III with subpopulation 1 only. Investigation on subpopulation 2 is stopped at stage 1, as the identification of heterogeneity of responses in favour of subpopulation 1 suggests that the response rate in subpopulation 2 is clinically insignificant as it is below its π_{02}. There is consequently no point in continuing the study with the entire population in phase III as would be performed with a standard Fleming design.

2.2.4. Adaptation of design parameters at stage 2. If only one subpopulation continues to stage 2 (situations C_1I_2, C_1I_2, I_1C_2 and I_1C_2), the number of patients recruited in stage 2 and the threshold b_2 have to be adapted to control α and β risks.

During the protocol phase, two separate Fleming designs should then also be planned: Independent Fleming design 1 (F_1) and Independent Fleming design 2 (F_2).

The following constraints are applied:

$$\alpha_{F1}=\alpha_{F2}=\alpha \quad \text{and} \quad \beta_{F1}=\beta_{F2}=\beta,$$

$$n_{1F1}=n_{11} \quad \text{and} \quad n_{1F2}=n_{21}.$$

The numbers of patients to be included at stage 2 if only one subpopulation continues to stage 2, n_{2F1} and n_{2F2}, and b_{2F1} and b_{2F2}, are therefore determined before the start of the trial.

2.2.5. Conclusions of the trial. At the end of stage 2, if only one subpopulation is investigated, only cumulative responses observed among $n_{i1}+n_{2Fi}$ patients are considered.

If both subpopulations continue to stage 2, n_{12} and n_{22} more patients enter the trial and the cumulative responses observed among $n_{11}+n_{21}+n_{12}+n_{22}$ patients are considered.

As previously, IP_{i2} are calculated and tested for heterogeneity of responses (Ψ_2).

The cumulative sample sizes at each stage and the final conclusions at the end of the two-stage trial are described in Tables II and III.

2.3. Determination of type I and type II errors

The decision to reject or not the null hypothesis is denoted by the binary variable Φ, where $\Phi=1$ corresponds to rejecting the null hypothesis and declaring the drug effective in at least one subpopulation

Table III. Decision rules at the end of the trial according to the investigated population during stage 2.

		R_d	
Investigated population during stage 2	Identification of an opposite treatment effect	$<b^*_{2d}$	$\geqslant b^*_{2d}$
Subpopulation 1 only		I_1I_2	E_1I_2
Entire population	$\Psi_2=1$	I_1I_2	E_1I_2
	$\Psi_2=0$	I_1I_2	E_1E_2
	$\Psi_2=2$	I_1I_2	I_1E_2
Subpopulation 2 only		I_1I_2	I_1E_2

$I_i=$ inefficacy of subpopulation i; $E_i=$ efficacy of subpopulation i; $R_d=$ cumulative number of responses.
$^*b_{2d}=b_2$ if both subpopulations continue in stage 2, $b_{2d}=b_{2Fi}$ if only subpopulation i continue in stage 2.

Table IV. Identification of type I and type II errors according to the decisions after stages 1 and 2.

		Decisions after stage 1								
		I_1I_2	C_1I_2	I_1C_2	C_1C_2	C_1I_2	I_1C_2	E_1E_2	E_1I_2	I_1E_2
Decisions after stage 2	I_1I_2				D_{12}					
	I_1		D_1			D_1				
	I_2			D_2			D_2			
	E_1E_2				B_{12}					
	E_1I_2				B_{12}					
	I_1E_2				B_{12}					
	E_1		B_1			B_1				
	E_2			B_2			B_2			
	Stop stage 1	C						A	A	A

$I=$ inefficacy (of subpopulation 1, I_1 or 2, I_2); $E=$ efficacy (of subpopulation 1, E_1 or 2, E_2); $C=$ continue to stage 2 (with subpopulation 1, E_1 or 2, E_2).
Probabilities A and B are calculated under π_{01} and π_{02}; Probabilities C and D are calculated under π_{11} and π_{12}.

and $\Phi=0$ corresponds to declaring the drug ineffective in both subpopulations. The decision Φ depends on the data observed in both subpopulations and at both stages. The prespecified type I and type II errors are denoted by α and β, respectively.

Thus, the design must satisfy: $P\{\Phi=1|H_{00}\}\leqslant\alpha$ and $P\{\Phi=0|H_{11}\}\leqslant\beta$. The power is the probability of declaring the drug effective in at least one subpopulation under H_{11}.

A, B, C and D are probabilities corresponding to the following situations (Table IV):

$$A=P(R_{11}+R_{21}\geqslant b_1|H_{00}),$$

$$B_{12}=P(R_{11}+R_{21}\in]a_1-b_1[\&\Psi_1=0|H_{00})\times P(R_{11}+R_{21}+R_{12}+R_{22}\geqslant b_2|H_{00}),$$

$$B_1=P(R_{11}+R_{21}<b_1\&\Psi_1=1|H_{00})\times P(R_{11}+R_{2F1}\geqslant b_{2F1}|H_{00}),$$

$$B_2=P(R_{11}+R_{21}<b_1\&\Psi_1=2|H_{00})\times P(R_{21}+R_{2F2}\geqslant b_{2F2}|H_{00}),$$

$$C=P(R_{11}+R_{21}\leqslant a_1\&\Psi_1=0|H_{11}),$$

$$D_{12}=P(R_{11}+R_{21}\in]a_1-b_1[\&\Psi_1=0|H_{11})\times P(R_{11}+R_{21}+R_{12}+R_{22}<b_2|H_{11}),$$

$$D_1=P(R_{11}+R_{21}<b_1\&\Psi_1=1|H_{11})\times P(R_{11}+R_{2F1}<b_{2F1}|H_{11}),$$

$$D_2=P(R_{11}+R_{21}<b_1\&\Psi_1=2|H_{11})\times P(R_{21}+R_{2F2}<b_{2F2}|H_{11}).$$

Then

$$P\{\Phi=1|H_{00}\}\leqslant\alpha\Leftrightarrow A+B_{12}+B_1+B_2\leqslant\alpha,$$

$$P\{\Phi=0|H_{11}\}\leqslant\beta\Leftrightarrow C+D_{12}+D_1+D_2\leqslant\beta.$$

2.4. Operating characteristics

We studied several scenarios of the true success rates in the two subpopulations (true success rates from 0 to 1, by 0.01), in order to study the operating characteristics of this adaptive stratified method.

Exact binomial probabilities are used to determine the maximal sample size to be included ($N\max = \max(\sum_{i=1}^{2}\sum_{s=1}^{2} n_{is}, \sum_{i=1}^{2} n_{i1}+n_2F_1, \sum_{i=1}^{2} n_{i1}+n_2F_2)$) and the expected number of patients under the null and alternative hypotheses ($E(n|H_{00})$, $E(n|H_{01})$, $E(n|H_{10})$, $E(n|H_{11})$). The true final conclusion rates, the probability of concluding Inefficacy or Efficacy on the whole population under H_{01} or H_{10}, the probability to detect heterogeneity of response between the two subpopulations at the first stage under H_{01}, H_{10} or H_{11} and the probability to continue to phase III under H_{00} or H_{11} are calculated.

Finally, these results were compared with those of a standard Fleming design, ie non-stratified and heterogeneous design, not taking into account the existence of two subpopulations.

First, we assume that $\pi_{01}=\pi_{02}$, $\Delta_1=\Delta_2$, $w=1$, $\alpha=0.05$, $\beta=0.1$ and $\gamma=0.6$. Two situations are further developed for $\pi_{0i}=0.25$: $\gamma=0.3$ and 0.8.

This method is illustrated by an example based on hypotheses used to conduct a closed but as yet unpublished trial, REMAGUS 02 [5]. The aim of this study was to determine the efficacy of two new treatments, Trastuzumab or Celecoxib, in combination with standard neoadjuvant chemotherapy for large operable and locally advanced breast cancer. The treatment plans were as follows:

- Patients who were HER_{2+} were randomized between Standard versus Standard+Trastuzumab.
- Patients who were HER_{2-} were randomized between Standard versus Standard+Celecoxib.

This trial was planned according to two parallel modified two-stage Fleming designs, one conducted in the HER_{2+} subpopulation and the other in the HER_{2-} subpopulation.

Trastuzumab and Celecoxib are considered to be new treatments (New) and the aim is to determine the efficacy of Standard +New in HER_{2+} and HER_{2-} subpopulations. Hence, for this example, patients who are randomized to Standard therapy alone are not considered. The published response rate to standard treatment is 15 per cent. In the context of this trial, New is considered to be potentially useful if it induces a response rate of at least 30 per cent in the HER_{2+} subpopulation and at least 25 per cent in the HER_{2-} subpopulation. The design parameters are as follows:

Study 1: HER_{2+} subpopulation: H_0: $\pi_1 \leqslant 0.15$&H_1: $\pi_1 \geqslant 0.30$, $\alpha=0.07$ and $\beta=0.10$.

Study 2: HER_{2-} subpopulation: H_0: $\pi_2 \leqslant 0.15$&H_1: $\pi_2 \geqslant 0.25$, $\alpha=0.09$ and $\beta=0.10$.

As about 25 per cent of women are HER_{2+}, w is equal to 3.

3. Results

Results for balanced design parameters (hypotheses and sample size) are presented in Tables V and VI. The maximal sample size is greater with the Adaptive Fleming design than with the standard heterogeneous design. However, expected sample sizes are very similar when response rates belong to the null hypothesis ($E(n|H_{00})$) or, on the contrary, to the alternative hypothesis ($E(n|H_{11})$). For example, when $\pi_{0i}=0.25$, $E(n|H_{00})$ is equal to 38.20 with the standard heterogeneous design and 38.47, 39.62 and 40.47 with the adaptive design for $\gamma=0.3$, 0.6 and 0.8, respectively. When only one subpopulation is sensitive, (($E(n|H_{01})$ or $E(n|H_{10})$), less than 10 per cent more patients are actually included with the adaptive Fleming design. One of the key advantages of this design is the number of true conclusions, leading to phase III trials conducted only on the true sensitive subpopulation. When only one subpopulation may benefit (H_{01} or H_{10}), the standard heterogeneous design cannot allow identification of this subpopulation, as heterogeneity cannot be determined, while the adaptive Fleming design allows determination of heterogeneity. True conclusions (Inefficacy in subpopulation 1 and Efficacy in subpopulation 2 under H_{01} or Efficacy in subpopulation 1 and Inefficacy in subpopulation 2 under H_{10}) are given in at least 10 per cent of trials and up to 26 per cent depending on the hypothesis. Moreover, the probability to conclude on inefficacy in both subpopulations when the true response rate is equal to π_{0i} in only one subpopulation is lower with the adaptive Fleming design than with the standard heterogeneous design. For example, under H_{01} or H_{10}, when 51.6 per cent of the trial's conclusions are 'Inefficacy in both arms' with the standard heterogeneous design, this rate is 38.1 per cent with the Adaptive design ($\pi_{0i}=0.2$). This means that fewer developments of promising drugs are stopped in phase II when one subpopulation is sensitive. For the same hypothesis, when 48.4 per cent of the trial's conclusions are 'Efficacy in both arms' with the standard heterogeneous design, this rate is 40.1

Table V. Maximal and expected sample sizes of the standard non-stratified heterogeneous Fleming two-stage design (H) and the stratified adaptive Fleming two-stage design (A), under null, alternative or combined hypothesis with equal response rates under the null hypothesis, $\Delta_i = 0.2$, $w = 1$, $\alpha = 0.05$ and $\beta = 0.1$.

		Maximal sample size		Expected sample size					
				A			H		
π_{0i}	γ	A	H	H_{00}^*	H_{01} or H_{10}	H_{11}	H_{00}	H_{01} or H_{10}	H_{11}
0.05	0.6	32	28	20.75	21.77	17.69	20.75	21.77	17.69
0.10	0.6	42	36	27.39	29.32	23.73	27.39	29.32	23.73
0.15	0.6	48	40	32.03	35.27	28.24	31.46	34.13	28.09
0.20	0.6	57	48	38.11	44.23	35.71	36.66	41.95	35.42
0.25	*0.3*	*62*	*52*	*38.47*	*45.27*	*37.85*	*38.20*	*44.41*	*37.79*
0.25	0.6	62	52	39.62	46.20	37.92	38.20	44.41	37.79
0.25	*0.8*	*62*	*52*	*40.47*	*47.79*	*38.25*	*38.20*	*44.41*	*37.79*
0.30	0.6	67	56	43.10	51.84	43.78	41.01	49.27	43.59
0.35	0.6	69	56	41.02	50.50	41.84	38.37	46.51	41.47
0.40	0.6	70	56	42.25	52.59	44.63	40.34	48.86	44.24
0.45	0.6	68	56	41.04	50.34	42.42	37.86	46.38	42.16
0.50	0.6	67	56	42.52	52.75	45.46	39.73	48.96	45.20
0.55	0.6	67	56	40.09	51.57	43.52	37.29	46.77	43.20
0.60	0.6	60	48	34.01	42.80	36.27	31.55	38.98	36.09
0.65	0.6	53	44	30.32	39.49	35.86	28.51	36.24	35.75
0.70	0.6	47	40	25.81	34.35	31.37	24.60	31.36	31.30
0.75	0.6	38	32	19.66	26.59	31.32	19.15	24.93	31.31

*Under H_{00}, $\pi_1 = \pi_{01}$ and $\pi_2 = \pi_{02}$; under H_{01}, $\pi_1 = \pi_{01}$ and $\pi_1 = \pi_{12}$; under H_{10}, $\pi_1 = \pi_{11}$ and $\pi_1 = \pi_{02}$; under H_{11}, $\pi_1 = \pi_{11}$ and $\pi_1 = \pi_{12}$.

per cent with the Adaptive design. This means that fewer non-sensitive patients enter phase III trials and fewer patients are exposed to toxic drugs, as the adaptive Fleming design detects the non-sensitive subpopulation at an earlier stage (up to 19.6 per cent depending on the hypothesis).

Determination of heterogeneity of response depends on γ values. The more the γ increases, the more the probability to detect the non-sensitive subpopulation at first stage increases (from 0.073 to 0.262). But the more the γ increases, the more type I error and the more the probability of not detecting a sensitive subpopulation at the first stage inflate. In the example where $\pi_{0i} = 0.25$ and for $\gamma = 0.3$, 0.6 and 0.8, the probabilities of continuing to phase III when no subpopulation could benefit (H_{00}) are 0.052, 0.055 and 0.064, respectively, while the initial α value was 0.05.

Even if the probability of not detecting a sensitive subpopulation while both are sensitive is not formally controlled by the design, it may be noted that this probability (calculated as a proportion of 'Efficacy in one subpopulation and Inefficacy in the other subpopulation' conclusion under H_{11}…) varies at first stage, from 0.006 to 0.042 (Table VI), and up to 0.04 when $\pi_{0i} = 0.25$ and $\gamma = 0.8$ for the final conclusion at the end of the trial (data not shown). Similar conclusions are reached when design parameters are unbalanced between the two subpopulations. As planned in the REMAGUS 02 Trial, 60 participants from subpopulation 1, i.e. with HER2+ receptors, were included, 30 at each stage. One hundred and ten participants from subpopulation 2, i.e. with HER2- receptors, were planned, 55 at each stage. As expected, maximum sample sizes with adaptive or standard heterogeneous designs were smaller than with two independent Fleming designs (Table VII). Expected sample sizes were very similar between standard heterogeneous and adaptive Fleming designs, except when only one subpopulation may benefit. Under H_{01} or H_{10}, true conclusion rates were improved with the adaptive Fleming design. As shown in Figure 2, the additional patients who enter stage 2 are those who benefit (Lower right and upper left quadrants where true response rates are greater than π_{0i} and 5–25 more patients on average are included due to population targeting) and who have been selected after stage 1 to continue during stage 2. Figure 3 illustrates the benefit in terms of false-negative conclusions.

4. Discussion

Subgroup analyses are often presented in study reports in the clinical literature. The most frequently reported reason for subgroup analysis is to show that treatment has a 'statistically significant' effect

Table VI. Operating characteristics of the standard non-stratified heterogeneous Fleming two-stage design (H) and the stratified adaptive Fleming two-stage design (A), under null, alternative or combined hypothesis with equal response rates under the null hypothesis, $\Delta i = 0.2$, $w = 1$, $\alpha = 0.05$ and $\beta = 0.1$.

		True final conclusion rates	Probability of concluding inefficacy on the whole population		Probability of concluding efficacy on the whole population		Probability of detecting the non-sensitive subpopulation at the first stage	Probability of not detecting a sensitive subpopulation at the first stage	Probability of continuing to Phase III	
		A	A	H	A	H	A	A	A	
π_{0i}	γ	H_{01}^{*} or H_{10}	H_{01} or H_{10}	H_{01} or H_{10}	H_{01} or H_{10}	H_{01} or H_{10}	H_{01} or H_{10}	H_{11}	H_{00}	H_{11}
0.05	0.6	0.000	0.361	0.361	0.639	0.639	0.000	0.000	0.060	0.941
0.10	0.6	0.012	0.521	0.521	0.467	0.479	0.000	0.000	0.042	0.899
0.15	0.6	0.143	0.458	0.557	0.399	0.443	0.148	0.020	0.052	0.893
0.20	0.6	0.217	0.381	0.516	0.401	0.484	0.217	0.030	0.063	0.926
0.25	*0.3*	*0.072*	*0.484*	*0.535*	*0.444*	*0.465*	*0.073*	*0.006*	*0.052*	*0.915*
0.25	*0.6*	*0.102*	*0.462*	*0.535*	*0.436*	*0.465*	*0.099*	*0.008*	*0.055*	*0.916*
0.25	*0.8*	*0.259*	*0.371*	*0.535*	*0.370*	*0.465*	*0.262*	*0.042*	*0.064*	*0.924*
0.30	0.6	0.135	0.427	0.519	0.438	0.481	0.128	0.010	0.061	0.929
0.35	0.6	0.160	0.423	0.542	0.416	0.458	0.166	0.017	0.063	0.919
0.40	0.6	0.195	0.423	0.558	0.382	0.442	0.196	0.024	0.062	0.919
0.45	0.6	0.135	0.468	0.580	0.397	0.420	0.138	0.010	0.058	0.910
0.50	0.6	0.167	0.465	0.598	0.368	0.402	0.167	0.013	0.054	0.914
0.55	0.6	0.186	0.465	0.622	0.349	0.378	0.194	0.015	0.048	0.915
0.60	0.6	0.122	0.500	0.612	0.378	0.388	0.127	0.006	0.056	0.910
0.65	0.6	0.110	0.591	0.699	0.298	0.301	0.116	0.004	0.037	0.883
0.70	0.6	0.107	0.616	0.724	0.276	0.276	0.111	0.002	0.034	0.893
0.75	0.6	0.074	0.681	0.755	0.245	0.245	0.076	0.000	0.027	0.911

*Under H_{00}, $\pi_1 = \pi_{01}$ and $\pi_2 = \pi_{02}$; under H_{01}, $\pi_1 = \pi_{01}$ and $\pi_1 = \pi_{12}$; under H_{10}, $\pi_1 = \pi_{11}$ and $\pi_1 = \pi_{02}$; under H_{11}, $\pi_1 = \pi_{11}$ and $\pi_1 = \pi_{12}$.

Table VII. Design parameters according to the two independent classical two-stage Fleming designs, the adaptive two-stage design (A) and a standard non-stratified heterogeneous Fleming two-stage design (H), based on REMAGUS02 trial hypothesis.

						Sample size			Thresholds			
						Stage 1	Stage 2		Stage 1	Stage 2		Maximal cumulated
Fleming design	Subpopulation	$\pi_{0i}-\pi_{1i}$	α	β	w^*	n_{i1}	n_{i2}	n_{2Fi}	a_1-b_1	b_2	b_{2Fi}	sample size
Double and independent (Remagus 02)	HER_{2+}	0.15–0.30	0.07	0.10	—	30	30	—	4–	14	—	170
	HER_{2-}	0.15–0.25	0.09	0.10		55	55	—	8–	22	—	
Adaptive	HER_{2+}	0.15–0.30	0.08	0.10	3	13	13	44	7–13	22	13	127
	HER_{2-}	0.15–0.25				39	39	75			23	
Heterogeneous	HER_{2+}&HER_{2-}	0.15–0.26	0.08	0.10	3	52	52	—	7–13	22	—	104

*Ratio $\frac{HER_{2-}}{HER_{2+}}$.

in one or more subgroups, although the trial shows no apparent overall effect [6–8]. These subgroups are usually not predefined before the trial. These types of subgroup analyses are usually descriptive and have no force of inference and do not provide any confirmatory evidence for subgroup treatment effects.

In this paper, we propose a new stratified adaptive phase II design based on the well-known two-stage Fleming design, allowing detection when only one of two subpopulations is sensitive rather than the whole population. Our method is based on identification of heterogeneity of responses; when the observed response rate is a prescribed amount less than π_{0i} in one subpopulation and a prescribed amount greater than π_{0i} in the other subpopulation, the method considers that there is heterogeneity of responses. Only the subpopulation with response rate above its π_{0i} is considered to present substantial evidence of efficacy to warrant further evaluation. The subpopulation with observed response rate below its π_{0i} is ignored because this observed response rate is considered as clinically insignificant. The procedure Ψ_s used to identify the target population decreases the false-negative rate. The more the γ increases, the more the heterogeneity is identified. The range of 0.6–0.8 is probably the only sensible values. Indeed small values like 0.3 lead to a need for a huge separation of the two observed response rates based on the biomarker that it seems unlikely. In our examples, both subpopulations present the same probability of response under the null hypothesis, but different π_{0i} are possible if there is a scientific rationale.

This new adaptive stratifed design differs from the two methods proposed by London and Chang [2] and Jones and Holmgren [3] in several aspects. First, it is based on a Fleming design, whereas these methods are based on a Simon design; hence, this adaptive design permits early stopping for inefficacy or efficacy. The primary motivation for the development of conditional and unconditional approaches by London and Chang [2] is to minimize the required sample size under given constrains on type I and type II error rates, while accounting for and permitting the response rates to differ by stratum. Stratification is used to improve the power requirement of a global test. In our method, stratification is first used to target the sensitive population. Identification of efficacy restricted to a subpopulation reduces the number of unethical trials in non-sensitive patients and allows continuation of drug development to phase III when it would previously have been stopped for the entire trial population.

Jones and Holmgren [3] have also proposed an adaptive Simon two-stage design that allows for preliminary determination of efficacy that may be restricted to a particular sub-population. According to the authors, when efficacy is restricted to one subpopulation, the power for making the correct decision to move forward in just this sub-group is quite comparable between their approach and two parallel Simon two-stage designs with approximately 9 per cent fewer subjects required. But the authors do not control the global type II error on the two parallel Simon designs. Moreover, their results show that the type I error is not always controlled at the nominal level.

Nevertheless, all these methods have similarities too. They all attempt to take into account heterogeneity of included populations. One trial replaces separate studies leading to save time and a large number of participants. The ratio between the sample sizes of the subpopulations is assumed to be known before the trial starts and is considered to be constant. This point should be well discussed with

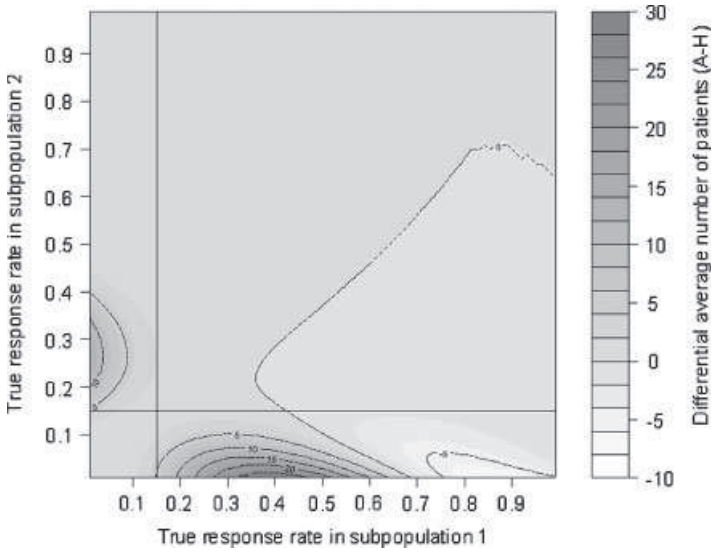

Figure 2. Differential average number of patients to be included between adaptive (A) and standard heterogeneous (H) designs according to true response rates in each subpopulation 1 and 2. Parameters are as follows: $\alpha=0.8$, $\beta=0.1$, $\gamma=0.6$, $w=3$, $\pi_{01}=\pi_{02}=0.15$, $\pi_{11}=0.3$, $\pi_{12}=0.25$.

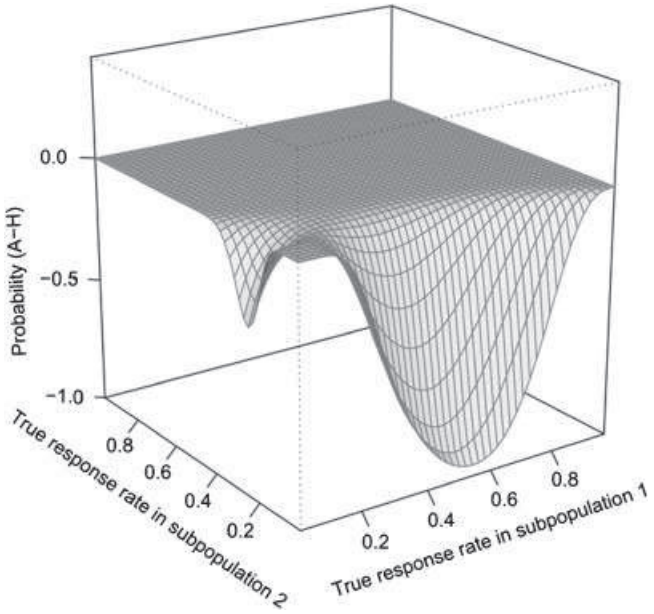

Figure 3. Differential probability to conclude of inefficacy in the whole population between adaptive (A) and standard heterogeneous (H) designs and according to true response rates in each subpopulation $I(i=1,2)$. Parameters are as follows: $\alpha=0.8$, $\beta=0.1$, $\gamma=0.6$, $w=3$, $\pi_{01}=\pi_{02}=0.15$, $\pi_{11}=0.3$, $\pi_{12}=0.25$.

clinicians in order to not slow patient recruitment.

It can be argued that this method is not randomized, which has been demonstrated to be the only way to distinguish prognosis and prediction [9–11]. Previous data showing that the marker is not prognostic or good historical data (stable over time) on response rate in both marker-defined groups separately would be helpful. Otherwise, it is impossible to determine if the marker is prognostic or predictive. The final outcome is to continue development of drug that could be of benefit to some patients, regardless of the reasons for treatment benefit and to propose phase III trials only to these patients.

In conclusion, this new stratified adaptive phase II design improves targeting of the populations entering phase III clinical trials. Fewer drug developments will be stopped in phase II due to treatment effect dilution and fewer non-sensitive patients will be exposed to toxic drugs. The power of Ψ_s may

be improved and is currently under investigation, together with incorporation of safety data in the statistical decision rules. An alternative to this approach is an approach-based hierarchical models, likely Bayesian, where there are assumed to be two response rates $p1$ and $p2$ in the two populations drawn from some overall population with a mean and variance [12].

Acknowledgements

We thank Dan Sargent for providing helpful comments during the preparation of this manuscript. The authors indicated no potential conflicts of interest.

References

1. Chen TT. Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine* 1997; **16**:2701–2711.
2. London WB, Chang MN. One- and two-stage designs for stratified phase II clinical trials. *Statistics in Medicine* 2005; **24**:2597–2611.
3. Jones CL, Holmgren E. An adaptive Simon Two-Stage Design for Phase 2 studies of targeted therapies. *Contemporary Clinical Trials* 2007; **28**:654–661.
4. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; **38**:143–151.
5. Pierga JY, Mathiot C, Extra JM, Tresca P, Asselah J, Sigal-Zafrani B, Brain E, Diéras V, Mignot L, Marty M. Circulating tumor celles detection in a randomized phase II trial of neoadjuvant chemotherapy for large operable and locally advanced breast cancer (REMAGUS 02): preliminary results. *Journal of Clinical Oncology* 2006; **24**:10637. ASCO Annual Meeting Proceedings Part I.
6. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *Journal of Clinical Epidemiology* 2004; **57**:229–236.
7. Grouin JM, Coste M, Lewis J. Subgroup analyses in randomized clinical trials: statistical and regulatory issues. *Journal of Biopharmaceutical Statistics* 2005; **15**:869–882.
8. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* 2002; **21**:2917–2930.
9. Ratain MJ, Humphrey RW, Gordon GB, Fyfe G, Adamson PC, Fleming TR, Stadler WM, Berry DA, Peck CC. Recommended changes to oncology clinical trial design: revolution or evolution? *European Journal of Cancer* 2008; **44**:8–11.
10. Ratain MJ, Sargent DJ. Optimising the design of phase II oncology trials: the importance of randomisation. *European Journal of Cancer* 2009; **45**:275–280.
11. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* 2005; **23**:2020–2027.
12. Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LH, Benjamin RS. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine* 2003; **22**:763–780.

Research Article

Statistics in Medicine

Received 29 January 2010, Accepted 7 January 2011 Published online 22 March 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4211

How a new stratified adaptive phase II design could improve targeting population

Caroline Tournoux-Facon,[a,c,d] Yann De Rycke[b*†] and Pascale Tubert-Bitter[c,d]

Phase II clinical trials in oncology are used for initial evaluation of the therapeutic efficacy of a new treatment regimen. Simon's two-stage or Fleming designs are commonly used for such trials. Treatment effect dilution may be observed because of heterogeneity of the population leading to negative conclusion on the whole population and termination of drug development while patients with particular characteristics may have benefit. Several authors have proposed alternative strategies based on Simon's design, including stratification on the patients' characteristics. In this paper, the authors develop a new stratified phase II design, allowing early stop of the trial for efficacy as it is an extension of Fleming's design. An example based on real data is given and the results from exact binomial calculations are developed to explore several assumptions on response rates, prevalence of each population and errors. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: adaptive phase II design; target population; stratification

1. Introduction

Phase III clinical trials often take years to recruit and adequately follow up patients. Even with a carefully planned phase II program, there may still be uncertainty at the beginning of phase III concerning the final target population. There is much interest, therefore, in being able to carry out adaptive phase II designs to target phase III patient population, i.e. the one who clearly may benefit from the new treatment. Indeed, situations exist where included patients characteristics are heterogeneous, leading to study failure (and treatment development stop) because of treatment effect dilution. One solution would be to conduct several phase II trials, one in each subpopulation of interest in order to decrease the probability of a false negative conclusion either in phase II or later in phase III trials.

In this context, the randomized phase II trial REMAGUS 02 [1] was planned at the Institut Curie in France. The aim of this study was to determine the efficacy of two new treatments, trastuzumab (H) or celecoxib (C), in combination with standard (S) neoadjuvant chemotherapy for large operable and locally advanced breast cancer. This trial was planned according to two parallel two-stage Fleming designs, one conducted in the HER_{2+} subpopulation and one in the HER_{2-} subpopulation. Women who were HER_{2+} were randomized between S ($n=58$) versus S+H ($n=62$), and women who were HER_{2-} were randomized between S ($n=108$) versus S+C ($n=112$). Sufficient efficacy of the new treatment has been demonstrated only in HER_{2+} women to warrant a phase III trial. Because of the

[a] *Centre d'Investigation Clinique P-0802, CHU Poitiers, France*
[b] *Service de Biostatistique, Institut Curie, Paris, France*
[c] *Inserm, CESP Centre for research in Epidemiology and Population Health, U1018, Biostatistics, F-94807, Villejuif, France*
[d] *Univ Paris-Sud, UMRS 1018, F-94807, Villejuif, France*
[*] *Correspondence to: Yann De Rycke, Service de Biostatistique, Institut Curie, Paris, France.*
[†] *E-mail: yann.de-rycke@curie.net*

very large number of women included in this phase II trial ($n=340$), clinicians asked the authors to develop an alternative method for future trials.

Stratified adaptive Simon two-stage designs have recently been proposed [2, 3]. However, these methods do not allow early termination in stage 1 for efficacy in one or several subgroups. The authors propose a new adaptive stratified phase II design based on the two-stage Fleming design [4].

The method assesses the heterogeneity of response between two subpopulations to identify if only one subpopulation present substantial efficacy to warrant further evaluation in stage 2 or in a phase III trial, leading to stop drug development for the other subpopulation. In a first approach [5], the heterogeneity assessment was based on comparing observed response rates to symmetric probability intervals around null response probabilities derived for both subpopulations as to meet prior probability coverage. In spite of attractive performances based on exact binomial probabilities, a lack of power may appear in some circumstances: when null response rates are close to 0 or 1, leading to large probability intervals, or when ratio between the two subpopulations differs from 1. It did not take into account large differences of treatment effect that may exist between the two subpopulations while only one observed response rate is not included in its probability interval. To account for the amount of difference between the two response rates rather than only their values compared to the null response rates and their intervals of probability, the authors propose an alternative approach leading to increased probability of true conclusion in case of heterogeneity. The structure of the paper is as follows. We present the new stratified adaptive phase II design, give the results of the application on REMAGUS 02 trial data, completed by exact binomial probabilities and discuss the method.

2. Methods

2.1. Notations and assumptions

We will assume that two subpopulations $i=1$ or 2 are predefined and that patients enter the clinical trial sequentially during stage $s=1$ or 2. The sample sizes for the first and second stages are denoted by n_{i1} and n_{i2} in each subpopulation i. We assume that the ratio between the two subpopulations may differ from 1 but is constant between the two stages and *a priori* defined: $n_{2s}=\omega\times n_{1s}$ (1).

The binary outcome for the jth patient in subpopulation i and stage s is denoted by X_{isj}, where $X_{isj}=1$ in the case of response and 0 in the case of failure. The cumulative number of responses within a given population at a given stage is denoted by $R_{is}=\sum_{j=1}^{n_{is}} X_{isj}$.

For this purpose, the probability of response in subpopulation i below which the investigated medicinal product is declared to be a low-activity drug is denoted by π_{0i}.

The null hypothesis is denoted by H_0: $\pi_1\leqslant\pi_{01}$ and $\pi_2\leqslant\pi_{02}$ and the alternative hypothesis is denoted by H_1: $\pi_1>\pi_{01}$ or $\pi_2>\pi_{02}$.

π_{1i} is defined as the lowest probability of success that we would like to detect for subpopulation i and $\Delta_i=\pi_{1i}-\pi_{0i}$. π_{01} may be different from π_{02} and Δ_1 may be different from Δ_2 as well.

2.2. General principles of the new design

Let consider a population of two groups of patients with one clinical or biological characteristic that may influence response to treatment. This characteristic may be a biomarker, previous treatments or comorbidity for example. The objective of this new stratified phase II design is to determine whether drug development in phase III should be continued on the whole population or stopped for a category of patients and continued with only one subpopulation if only one subpopulation seems sensitive to the treatment. For this purpose, we developed a method in which second-stage sample size and decision rules depend on the observed response rates at the first stage and the identification of heterogeneity of responses.

To introduce the procedure for the heterogeneity assessment between the two subpopulations, we consider the statistics $d_s=\sum_{i=1}^{2}|d_{is}|$ and $S_s=\sum_{i=1}^{2}\text{sign}(d_{is})$, $s=1,2$, where sign(x) is equal to 1, 0 or -1 according to x being greater than 0, equal to 0 or less than 0, and $d_{is}=(R_{is}/n_{is})-\pi_{0i}$ (difference between observed response rate and probability of response under H_0).

The heterogeneity test at γ level under H_0 has the form $T_s=\text{Ind}\{d_s>c_s \text{ and } S_s=0\}$, where c_s is a function of γ; $c_s=\min(c|P(d_s>c \text{ and } S_s=0|H_0)\leqslant\gamma)$.

Statistics in Medicine

Table I. Decision rules at the end of first stage, after inclusion of $n_{11}+n_{21}$ patients.

	$R_{11}+R_{21}\leqslant a_1$	$R_{11}+R_{21}\in]a_1-b_1[$	$R_{11}+R_{21}\geqslant b_1$
Ψ_1=1	C_1I_2	C_1I_2	E_1I_2
Ψ_1=0	I_1I_2	C_1C_2	E_1E_2
Ψ_1=2	I_1C_2	I_1C_2	I_1E_2

I = Inefficacy (of subpopulation 1, I_1 or 2, I_2); E = Efficacy (of subpopulation 1, E_1 or 2, E_2); C = Continue to stage 2 (with subpopulation 1, C_1 or 2, C_2); I, E and C are binary variables.

Table II. Decision rules at the end of the trial according to the investigated population during stage 2.

Investigated population during stage 2	Identification of an opposite treatment effect	Cumulated number of responses $<b_2$ §	Cumulated number of responses $\geqslant b_2^{\S}$
Subpopulation 1		I_1I_2	E_1I_2
Entire population	$\Psi_2=1$	I_1I_2	E_1I_2
	$\Psi_2=0$	I_1I_2	E_1E_2
	$\Psi_2=2$	I_1I_2	I_1E_2
Subpopulation 2		I_1I_2	I_1E_2

I = Inefficacy (of subpopulation 1, I_1 or 2, I_2); E = Efficacy (of subpopulation 1, E_1 or 2, E_2); § $b_2=b_{2F1}$ (b_{2F2}) if only subpopulation 1 (2) continue in stage 2.

The determination of heterogeneity and identification of potential treatment activity in favour of population i is denoted Ψ_s, with:

$$\Psi_s=0 \quad \text{when } T_s=0$$

$$\Psi_s=i \quad \text{when } T_s=1 \quad \text{and} \quad d_{is}>0$$

At the end of first stage, information from Ψ_1 is combined with the total number of responses $R_{11}+R_{21}$ leading to conclusions described in Table I.

If only one subpopulation continues to stage 2 (situations C_1I_2, C_1I_2, I_1C_2 and I_1C_2), the number of patients recruited in stage 2 and the threshold b_2 have to be adapted to control α and β risks. During the protocol phase, two separate Fleming designs should then also be planned: Independent Fleming design 1 (F_1) and Independent Fleming design 2 (F_2).

The following constraints are applied:

$$\alpha_{F1}=\alpha_{F2}=\alpha$$

$$\beta_{F1}=\beta_{F2}=\beta$$

$$n_{1F1}=n_{11} \quad \text{and} \quad n_{1F2}=n_{21}$$

The numbers of patients to be included at stage 2 if only one subpopulation continues to stage 2, n_{2F1} and n_{2F2}, and b_{2F1} and b_{2F2}, are therefore determined before the start of the trial.

If the two subpopulations continue in stage 2, as previously, d_2 and c_2 are combined to Ψ_2 in order to conclude to either inefficacy or efficacy of the new therapy on the whole population (Table II). Determination of α and β errors is presented in Appendix A.

2.3. *Operating characteristics*

2.3.1. Application of the new design on REMAGUS 02 trial data. For this purpose, data from S+H and S+C arms are pooled to determine the efficacy of Standard (S) and New treatments (H or C) combination in HER_{2+} and HER_{2-} subpopulations. Patients randomized to Standard arms are not considered in this paper. According to the protocol, response rate with standard treatment is stated at 15 per cent from the literature and around one-third of women is HER_{2+}. Combinations are considered of interest if response rates are at least 30 per cent in the HER_{2+} subpopulation and 25 per cent in the HER_{2-} subpopulation.

The design parameters are stated as follows:

- Subpopulation 1 HER_{2+}: H_0: $\pi_1\leqslant 0.15$ & H_1: $\pi_1\geqslant 0.30$.
- Subpopulation 2 HER_{2-}: H_0: $\pi_2\leqslant 0.15$ & H_1: $\pi_2\geqslant 0.25$.

Table III. Decision rules at the end of the trial according to the investigated population during stage 2.

Fleming design		Standard	Adaptive		Double	
Population (P)		P_1: 33 per cent P_2: 66 per cent	P_1: 33 per cent	P_2: 66 per cent	P_1: 100 per cent	P_2: 100 per cent
Hypothesis		π_0: 0.15 π_1: 0.26 α: 0.05 β: 0.10	π_0: 0.15 π_1: 0.30 α: 0.05 β: 0.10 γ: 0.6, $c_1=0.14$, $c_2=0.10$	π_0: 0.15 π_1: 0.25	π_0: 0.15 π_1: 0.30 α: 0.05 β: 0.10	π_0: 0.15 π_1: 0.25 α: 0.05 β: 0.10
Sample sizes	Stage 1 Stage 2	54 (18+36) 54 (18+36)	18 18 or 46	36 36 or 100	32 32	68 68
Stage 1	Observed responses Decision	$10(5+5)\in]a_1-b_1[$ Continue stage 2	$5+5\in]a_1-b_1[, d_1=0.14$ $\Psi_1=1$ (Continue subpopulation 1 only)		$9\in]a_1-b_1[$ Continue stage 2	$10\in]a_1-b_1[$ continue stage 2
Stage 2	Observed responses	21 $(11+10)<b_2$	$16\geqslant b_{2F1}$	—	$16\geqslant b_2$	$17<b_2$
Final conclusion		Inefficacy 1 and 2	Efficacy 1	Inefficacy 2	Efficacy 1	Inefficacy 2
cumulated included sample size		54+54=108	18+36+46=100		32+32+68+68=200	

In the REMAGUS 02 protocol, α and β risks were 0.07 and 0.10 in the HER_{2+} group and 0.09 and 0.10 in the HER_{2-} group for sample size considerations. For this purpose, α risk is stated at 0.05 and β risk is stated at 0.10, leading to the necessity to simulate supplementary patients and responses. Observed response rates were used for binomial calculation. Final conclusions and cumulated included sample sizes are compared between:

- The new adaptive design ('Adaptive Design').
- A standard Fleming design, ie non-stratified and heterogeneous design not taking into account the existence of two subpopulations but for this purpose, taking into account the ratio between the prevalence of the two subpopulations ('Standard Design').
- The REMAGUS 02 trial ('Double design'), including additional simulated responses.

Exact binomial probabilities

To further describe operating characteristics, results from this real example are completed by exact binomial probabilities to compare Standard and Adaptive designs. We assume that $\pi_{01}=\pi_{01}$, $\Delta_1=\Delta_1=0.2$, $w=1$, $\alpha=0.05$, $\beta=0.1$ and $\gamma=0.6$. We studied true success rates from 0 to 1, by 0.01 to determine the maximal sample size to be included

$$\text{Nmax}=\max\left(\sum_{i=1}^{2}\sum_{s=1}^{2}n_{is},\ \sum_{i=1}^{2}n_{i1}+n_{2F1},\ \sum_{i=1}^{2}n_{i1}+n_{2F2}\right)$$

the expected number of patients under the null and alternative hypotheses ($E(n|H_{00})$, $E(n|H_{01})$, $E(n|H_{10})$, $E(n|H_{11})$), the true final conclusion rates, the probability of concluding inefficacy or efficacy on the whole population under H_{01} or H_{10}, the probability to detect a non-sensitive subpopulation at the first stage under H_{01} or H_{10} and the probability to continue to phase III under H_{00}.

3. Results

Results inspired from the REMAGUS 02 trial are presented in Table III.

3.1. Double design

Because $\alpha=0.05$, two additional responses are simulated in the HER_{2+} group and 28 additional responses in the HER_{2-} group. A total of 200 patients are evaluated: 64 HER_{2+} women in the first Fleming design and 136 HER_{2-} women in the second Fleming design. Efficacy of the new therapy is demonstrated only for the first population, i.e. HER_{2+} patients.

3.2. Standard design

A standard heterogeneous and non-stratified Fleming design includes a total of 108 patients (36 HER_{2+} women and 72 HER_{2-} women) and the trial is negative: New therapies are considered as ineffective. Treatment effect dilution leads to stop drugs development for the entire population.

3.3. Adaptive design

With the new stratified adaptive Fleming method, the cumulated included sample size is 100: 64 HER_{2+} women and 36 HER_{2-} women. The conclusion is as follows: efficacy of the new combined therapy for HER_{2+} women and inefficacy of the new combined therapy for the HER_{2-} women. As many as HER_{2+} women have been included in the new adaptive stratified design compared to the Double Fleming design but 100 less HER_{2-} women have participated. The new adaptive stratified design leads to the right conclusion with fewer patients, in particular from the subpopulation that does not benefit because heterogeneity of response between the two subpopulations has been detected early, at the first stage.

Results from exact binomial probability calculations confirm these encouraging conclusions (Table IV). Indeed, expected sample sizes are very similar between standard and adaptive designs, in particular when response rates are closed to the null hypothesis ($E(n|H_{00})$) or the alternative hypothesis ($E(n|H_{11})$) in the two subpopulations. Under H_{01} or H_{10}, until 20 per cent more patients may be included with adaptive stratified design, but are in fact those whose could benefit. Indeed under H_{01} or H_{10},

Table IV. Operating characteristics of two-stage designs using the standard (heterogeneous and non stratified) two-stage design and the adaptive two-stage design, under null and alternative hypothesis.

	Maximal sample size		Expected sample size						True final conclusion rates	Probability of concluding inefficacy on the whole population		Probability of concluding efficacy on the whole population		Probability of detecting the non sensitive subpopulation at 1st stage	Probability to continue to phase III
			Adaptive			Standard			Adaptive	Adaptive	Standard	Adaptive	Standard	Adaptive	Adaptive
π_{0i}	Adaptive	Standard	H_{00}	H_{01} or H_{10}	H_{11}	H_{00}	H_{01} or H_{10}	H_{11}	H_{01} or H_{10}	H_{01} or H_{10}	H_{01} or H_{10}	H_{01} or H_{10}	H_{01} or H_{10}	H_{01} or H_{10}	H_{00}
0.05	32	28	21.36	26.14	21.57	21.11	24.77	21.05	0.415	0.306	0.381	0.277	0.619	0.394	0.052
0.10	42	36	28.83	33.96	27.74	27.78	32.09	27.36	0.324	0.373	0.562	0.301	0.438	0.321	0.045
0.15	48	40	32.63	38.87	32.47	31.78	36.39	31.78	0.316	0.408	0.582	0.275	0.418	0.320	0.048
0.20	57	48	37.72	44.53	35.96	36.66	41.95	35.42	0.286	0.391	0.516	0.322	0.484	0.267	0.059
0.25	62	52	40.70	50.40	42.63	38.46	46.43	41.80	0.333	0.383	0.545	0.283	0.455	0.313	0.056
0.30	67	56	44.38	54.32	44.58	41.01	49.27	43.59	0.365	0.341	0.519	0.292	0.481	0.346	0.068
0.35	70	56	43.39	54.64	46.48	38.62	48.33	45.59	0.311	0.367	0.549	0.321	0.451	0.292	0.066
0.40	70	56	44.93	55.02	45.23	40.34	48.86	44.24	0.346	0.361	0.558	0.292	0.442	0.323	0.068
0.45	68	56	42.95	54.71	47.48	38.08	48.05	46.41	0.347	0.373	0.588	0.278	0.412	0.347	0.062
0.50	67	56	44.16	54.41	45.78	39.73	48.96	45.20	0.289	0.394	0.598	0.316	0.402	0.288	0.061
0.55	67	56	42.80	54.96	48.47	37.46	48.02	47.79	0.325	0.400	0.629	0.274	0.371	0.318	0.050
0.60	60	48	35.59	46.63	41.54	31.79	40.57	40.80	0.305	0.416	0.622	0.278	0.378	0.298	0.057
0.65	53	44	31.37	41.85	40.57	28.63	37.18	40.18	0.256	0.504	0.706	0.240	0.294	0.256	0.040
0.70	47	40	29.23	38.83	37.08	24.73	32.42	36.70	0.275	0.472	0.734	0.253	0.266	0.289	0.043
0.75	38	32	21.91	30.33	31.49	19.15	24.93	31.31	0.237	0.518	0.755	0.245	0.245	0.246	0.037

Under H_{00}, $\pi_1=\pi_{01}$ & $\pi_2=\pi_{02}$; Under H_{01}, $\pi_1=\pi_{01}$ & $\pi_2=\pi_{12}$; Under H_{10}, $\pi_1=\pi_{11}$ & $\pi_2=\pi_{02}$; Under H_{11}, $\pi_1=\pi_{11}$ & $\pi_2=\pi_{12}$.

the true conclusions rate, leading to phase III trials conducted only on the true sensitive subpopulation is ranged from 23.7 to 41.5 per cent. Moreover, under H_{01} or H_{10} the probabilities to conclude 'Inefficacy in both arms' and 'Efficacy in both arms' are 18.9 and 12.7 per cent less with the adaptive design, compared to the standard heterogeneous and non-stratified design. This means that less development of promising drugs are stopped in phase II and that less non-sensitive patients enter in phase III trials. The adaptive design identifies early the non-sensitive subpopulation (probability to detect the non sensitive subpopulation at the first stage, median [minimum–maximum]: 0.31 [0.25–0.35]). However, due to the small number of patients included at first stage, the probability to conclude inefficacy in one subpopulation and efficacy in the other one at the first stage is very small, less than 5 per cent on these examples (data not shown). The probability to continue in phase III when no subpopulation could benefit (H_{00}) ranges from 0.037 to 0.068; the type I error is well controlled (initial α value stated at 0.05).

4. Discussion

Pocock reviewed 50 trials published in major journals in 1997 and noted that 70 per cent reported a median of four subgroup analyses [6]. The number of subgroup analyses presented in study reports reflects the need for improving targeting population entering in phase III trials.

In this paper we propose an adaptive stratified phase II design based on the two-stage Fleming design. It allows for detecting between two pre-specified subpopulations, the sensitive one when only one could benefit. Prevalence of the characteristic separating the two subpopulations should be well known before the trial starts in order to not slow patient recruitment and even when the endpoint can be assessed in a relative short period of time. Our method is based on the determination of the heterogeneity in the responses of trial subpopulations. It takes into account the amount of response difference between the two subpopulations and position around their respective π_{0i}. In case of heterogeneity, only one targeted subpopulation is considered as having substantial evidence of sufficient efficacy to warrant further evaluation. With this method, drug development is less often stopped for the entire phase II population and less non sensitive patients are then exposed to toxic drugs in the second part of phase II trials, and next in phase III trials. This method is not randomized which has been demonstrated to be the only way to distinguish prognosis and prediction [7, 8]. Moreover, such randomized designs reduce the potential of bias, existent in comparisons with historical controls, but they also substantially increase the sample size requirements and are not always applicable in all situations [9]. An extension to a randomized design and comparison with randomized biomarker-adaptive designs [10, 11] is under process.

The adaptive method is simple to implement and an R package will be available soon.

Appendix A: Determination of type I and II errors

The decision to reject the null hypothesis or not is denoted by the binary variable Φ, where $\Phi=1$ corresponds to rejecting the null hypothesis and declaring the drug effective in at least one subpopulation or $\Phi=0$ where we declare the drug ineffective in both subpopulations. The decision Φ depends on the data that are observed from both subpopulations and both stages. The pre-specified type I and type II errors will be denoted by α and β, respectively; thus, design must satisfy:

$$P\{\Phi=1|\ \pi_1=\pi_{01} \text{ and } \pi_2=\pi_{02}\}\leqslant\alpha \quad \text{and} \quad P\{\Phi=0|\pi_1=\pi_{11} \text{ and } \pi_2=\pi_{12}\}\leqslant\beta$$

A, B, C and D are probabilities corresponding to the following situations:

$$A = P(R_{11}+R_{21}\geqslant b_1 \& \Psi_1=0|H_0)+P(R_{11}+R_{21}\geqslant b_1 \& \Psi_1=1|H_0)+P(R_{11}+R_{21}\geqslant b_1 \& \Psi_1=2|H_0)$$

$$\begin{aligned} B_{12} = {} & P(R_{11}+R_{21}\in]a_1-b_1[\& \Psi_1=0|H_0)\times P(R_{11}+R_{21}+R_{12}+R_{22}\geqslant b_2 \& \Psi_2=0|H_0) \\ & +P(R_{11}+R_{21}\in]a_1-b_1[\& \Psi_1=0|H_0)\times P(R_{11}+R_{21}+R_{12}+R_{22}\geqslant b_2 \& \Psi_2=1|H_0) \\ & +P(R_{11}+R_{21}\in]a_1-b_1[\& \Psi_1=0|H_0)\times P(R_{11}+R_{21}+R_{12}+R_{22}\geqslant b_2 \& \Psi_2=2|H_0) \end{aligned}$$

$$\begin{aligned} B_1 = {} & P(R_{11}+R_{21}\leqslant a_1 \& \Psi_1=1|H_0)\times P(R_{11}+R_{2F1}\geqslant b_{2F_1}|H_0) \\ & +P(R_{11}+R_{21}\in]a_1-b_1[\& \Psi_1=1|H_0)\times P(R_{11}+R_{2F1}\geqslant b_{2F1}|H_0) \end{aligned}$$

$$B_2 = P(R_{11}+R_{21}\leqslant a_1 \& \Psi_1 = 2|H_0) \times P(R_{21}+R_{2F2}\geqslant b_{2F2}|H_0)$$
$$+P(R_{11}+R_{21}\in]a_1-b_1[\& \Psi_1 = 2|H_0) \times P(R_{21}+R_{2F2}\geqslant b_{2F2}|H_0)$$
$$C = P(R_{11}+R_{21}\leqslant a_1 \& \Psi_1 = 0|H_1)$$
$$D_{12} = P(R_{11}+R_{21}\in]a_1-b_1[\& \Psi_1 = 0|H_1) \times P(R_{11}+R_{21}+R_{12}+R_{22}<b_2|H_1)$$
$$D_1 = P(R_{11}+R_{21}\leqslant a_1 \& \Psi_1 = 1|H_1) \times P(R_{11}+R_{2F1}<b_{2F1}|H_1)$$
$$+P(R_{11}+R_{21}\in]a_1-b_1[\& \Psi_1 = 1|H_1) \times P(R_{11}+R_{2F1}<b_{2F1}|H_1)$$
$$D_2 = P(R_{11}+R_{21}\leqslant a_1 \& \Psi_1 = 2|H_1) \times P(R_{21}+R_{2F2}\geqslant b_{2F2}|H_1)$$
$$+P(R_{11}+R_{21}\in]a_1-b_1[\& \Psi_1 = 2|H_1) \times P(R_{21}+R_{2F2}<b_{2F2}|H_1)$$

Then:

$$P\{\Phi = 1|\pi_1 = \pi_{01} \text{ and } \pi_2 = \pi_{02}\} \leqslant \alpha \text{ is equal to } A+B_{12}+B_1+B_2 \leqslant \alpha$$
$$P\{\Phi = 0|\pi_1 = \pi_{11} \text{ and } \pi_2 = \pi_{12}\} \leqslant \beta \text{ is equal to } C+D_{12}+D_1+D_2 \leqslant \beta$$

References

1. Pierga JY, Delaloge S. Espié M, Brain E, Sigal-Zafrani B, Mathieu MC, Bertheau P, Guinebretière JM, Spielmann M, Savignoni A, Marty M. A multicenter randomized phase II study of sequential epirubicin/cyclophosphamide followed by docetaxel with or without celecoxib or trastuzumab according to HER2 status, as primary chemotherapy for localized invasive breast cancer patient. *Breast Cancer Research and Treatment* 2010; **122**(2):429–437.
2. Jones CL, Holmgren E. An adaptive Simon two-stage design for phase 2 studies of targeted therapies. *Contemporary Clinical Trials* 2007; **28**(5):654–661.
3. London WB, Chang MN. One- and two-stage designs for stratified phase II clinical trials. *Statistics in Medicine* 2005; **24**(17):2597–2611.
4. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; **38**(1):143–151.
5. Tournoux-Facon C, De Rycke Y, Tubert-Bitter P. Targeting population entering phase III trials: a new stratified adaptive phase II design. *Statistics in Medicine* 2011; DOI: 10.1002/sim.4148.
6. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* 2002; **21**(19):2917–2930.
7. Ratain MJ, Sargent DJ. Optimising the design of phase II oncology trials: the importance of randomisation. *European Journal of Cancer* 2009; **45**:275–280.
8. Sargent DJ, Conley BA, Allegra C, Collete L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* 2005; **23**:2020–2027.
9. Rubinstein L, Crowley J. Randomized phase II designs. *Clinical Cancer Research* 2009; **15**(6):1883–1890.
10. Wang SJ, O'Neill RT, Hung HJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics* 2007; **6**:227–244.
11. Zhou X, Liu S, Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine. *Clinical Trials* 2008; **5**:181–193.