Health-Related Quality of Life and Statistical Challenges in Trial-Based Economic Evaluations

Jornada Ben, Angela

2023

DOI (link to publisher)
10.5463/thesis.104

document version
Publisher's PDF, also known as Version of record

Health-Related Quality of Life and Statistical Challenges in Trial-Based Economic Evaluations

Ângela Jornada Ben

Health-Related Quality of Life and Statistical Challenges in Trial-Based Economic Evaluations

Ângela Jornada Ben

Colophon

The studies presented in this thesis were conducted at the Department of Health Sciences of the Vrije Universiteit Amsterdam. This PhD thesis was embedded within the Faculty of Science, Vrije Universiteit Amsterdam and Amsterdam Public Health research institute.

The studies in this thesis were funded by the Amsterdam Public Health Research Institute, the EuroQol Research Foundation, the Van Meenen middelen, ZonMw, and São Paulo Research Foundation (FAPESP).

English title: Health-Related Quality of Life and Statistical Challenges in Trial-Based Economic Evaluations.

Dutch title: Gezondheidsgerelateerde Kwaliteit van Leven en Statistische Uitdagingen in Empirische Economische Evaluaties.

ISBN: 978-90-833062-3-0
Lay-out: Legatron Eletronic Publishing
Cover: https://huggingface.co/spaces/stabilityai/stable-diffusion
Printed by: Proefschrift Specialist

VRIJE UNIVERSITEIT

HEALTH-RELATED QUALITY OF LIFE AND STATISTICAL CHALLENGES IN TRIAL-BASED ECONOMIC EVALUATIONS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op woensdag 15 februari 2023 om 11.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Ângela Jornada Ben

geboren te Passo Fundo, Brazilië

promotor:	prof.dr. J.E. Bosmans
copromotoren:	dr. J.M. van Dongen dr. M. El Alili
promotiecommissie:	prof.dr. A. de Wit prof.dr. W. Janssens dr. E. Stolk prof.dr. J. van Busschbach prof.dr. H. Koffijberg

Contents

CHAPTER 1

General introduction

Economic evaluations

Healthcare systems worldwide are struggling with the introduction of new – and oftentimes more expensive – health technologies (e.g., medicines, medical procedures, and screening programs) due to a shortage of healthcare resources.[1–3] In the Netherlands, for example, The increasing introduction and use of novel health technologies in combination with the ageing of the population are expected to increase healthcare spending by 2.9% per year up until 2040.[4–7] As a consequence, finding ways to allocate already limited healthcare resources as efficiently as possible has become a high priority in many countries.[8–10]

Economic evaluations can help inform resource allocation decisions in healthcare by providing insight into the "value for money" of health technologies.[11] In an economic evaluation, the difference in costs between two or more health technologies is related to the difference in effects between the technologies.[12] By doing so, economic evaluations provide an estimate of the incremental cost per unit of effect gained, also known as the "Incremental cost-effectiveness ratio" (ICER). A health technology can be considered cost-effective if the ICER is lower than the maximum amount of money that healthcare decision-makers are willing to pay per unit of effect gained, or if a health technology turns out to be both less expensive and more effective compared to its comparator.

Over the last decades, more and more healthcare systems worldwide have been reformed to consider evidence of the cost-effectiveness of health technologies in their decision-making processes.[13–15] In the Netherlands, for example, cost-effectiveness evidence is one of the mandatory criteria established by the Dutch Health Technology Assessment (HTA) agency (i.e., the National Health Care Institute) when advising the Minister of Health, Welfare, and Sports about the inclusion of pharmaceutical products, but very limited when dealing with non-pharmaceutical health technologies, in the basic health insurance package.[13–16] In this way, grant organizations within and outside the Netherlands increasingly require that economic evaluations are conducted alongside the clinical trials that they fund (e.g., the National Institute for Health and Care Research – NIHR in the UK, ZonMw in the Netherlands).[17,18]

Different types of economic evaluations exist, which differ in how effect outcomes are measured and valued.[11,12] When the outcomes of the compared health technologies are expressed as a clinical effect (e.g., pain intensity), an economic evaluation is labelled as a **cost-effectiveness analysis (CEA)**. When using clinical effects as outcomes in an economic evaluation, the results can only be compared to results from studies on health technologies that aim to improve that specific clinical effect. However, healthcare decision-makers often need to make a choice between health technologies for different disorders that thus target different clinical effects. Therefore, more generic outcome measures are needed. The most well-known example of such a comprehensive outcome measure is the Quality-Adjusted Life-Year (QALY). If the QALY is used as an outcome in an economic evaluation, it is labelled as a **cost-utility analysis (CUA)**. Both CEA and CUA provide information on how to maximize health benefits within a specific budget.[12] A **cost-benefit analysis (CBA)** places a monetary value on an effect outcome (e.g., by converting sickness absence days in costs) and informs healthcare decision-makers about whether or not a health technology generates financial savings compared with control.[12] Finally, in a **cost-minimization analysis** (**CMA**) the outcomes of two health technologies are considered to be equivalent and only costs are compared.[12]

An economic evaluation can be designed as a model-based economic evaluation or a trial-based economic evaluation.[19,20] Model-based economic evaluations use a decision-analytic model to estimate the differences in costs and effects between two or more health technologies and are parametrized using multiple sources of information from the literature (e.g., clinical trials, cohorts, systematic reviews, and metanalysis), registries, electronic health records, and other available sources as input.[19] In this thesis, we focus on trial-based economic evaluations, which are economic evaluations that are conducted alongside clinical trials and are sometimes referred to as "piggy-back" studies. The random allocation of patients across study conditions in most trial-based economic evaluations increases the internal validity of such studies, while the prospective collection of patient-level cost and effect data reduces the possible influence of information bias (i.e., biases that arise from systematic differences in the data collection).[21–23] Preferably, a trial-based economic evaluation is designed as a pragmatic or naturalistic trial, meaning that it resembles daily practice as much as possible. In this way, findings can be easily generalized to real-world settings (i.e., increasing external validity).

Health-related quality of life, utilities, and QALYs

HTA agencies typically require researchers to use QALYs as the primary outcome in their economic evaluation, because QALYs enable the comparison of a broad range of health technologies for different health conditions.[11,13,14,24]

QALYs combine both the length of life and health-related quality of life (HRQoL) in a single index.[25] The length of life is defined as the amount of time an individual experiences a particular health state (e.g., 3 years).[26] HRQoL is expressed as a utility value, typically representing the general public's preference for a specific health state.[26] Utility values are anchored at 0 and 1. A zero indicates that it is valued as being equal to "death", while a utility value of 1 indicates that a health state is valued as being equal to "full health". Negative utility values can also occur and indicate that a health state is valued as being "worse than death".[26]

One of the most commonly used questionnaires for assessing HRQoL and estimating utility values in trial-based economic evaluations is the EQ-5D.[27,28] The descriptive system of the EQ-5D asks individuals to describe their health state based on five health dimensions (i.e., mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) with 3 or 5 response levels per health dimension (i.e., EQ-5D-3L and EQ-5D-5L, respectively).[29,30] The EQ-5D version with 3-levels describes 243 health states and was recently extended to a 5-level version to improve the sensitivity and responsiveness of the instrument, resulting in a total of 3,125 health states.[31] The health states described by the two EQ-5D versions can be converted in utility values using so-called EQ-5D value sets or tariffs.[32] Such value sets are ideally obtained from a country population sample using valuation protocols developed by the EuroQol group, which have evolved over the years to improve the validity and reliability of utility values derived from the general public.[33] This is typically done per country to account for populations' sociocultural differences.

The importance of doing it "right"

An important prerequisite for using trial-based economic evaluation results in healthcare decision-making is that they are valid and reliable. Amongst others, this means that research should be "*scrupulous*". Scrupulousness means that research is conducted "*using methods that are scientific or scholarly and exercising the best possible care in designing, undertaking, reporting and disseminating research*".[34] Scrupulousness is one of the research integrity principles stated in the Dutch Code of Conduct for Research Integrity[34] which was solidly built from the Nuremberg Code[35] and the Singapore Statement.[36] Scrupulousness also implies that the use of less-than-optimal methods when conducting trial-based economic evaluations can result in misleading conclusions and a waste of already scarce healthcare resources. Nonetheless, research indicates that the methodological quality of trial-based economic evaluations is far from optimal.[37,38]

Existing gaps in knowledge

In the area of trial-based economic evaluations, various gaps in knowledge exist that make it unclear how certain design and analysis steps can be scrupulously conducted. Three of these gaps in knowledge will be addressed in this thesis and will be discussed in greater detail in the next sections.

The impact of using crosswalks on healthcare decision-making

The two versions of the EQ-5D, i.e., the EQ-5D-3L and EQ-5D-5L, are not equivalent; neither in terms of the number of response levels and wording nor in terms of their country-specific valuation protocols.[31,33] With the increasing uptake of the EQ-5D-5L by HTA agencies, academic groups, and companies, valuation studies of the EQ-5D-5L are being conducted in many countries. To guarantee that the resulting value sets are valid and reliable, researchers conducting EQ-5D valuation studies require extensive training and are recommended to use standardized EQ-5D valuation methods that are evolving over time.[33] However, conducting valuation studies is time-consuming, which explains why EQ-5D-5L value sets are not available yet for many countries. As an interim solution, mapping approaches, where EQ-5D-5L responses are mapped onto EQ-5D-3L responses and vice versa, were developed to estimate utility values when value sets are missing for a certain version of the EQ-5D or country.[39–41]

Given that HTA agencies might be confronted with evidence that is based on different EQ-5D versions and/or value sets, guidance on choosing the most appropriate utility scoring method is needed to ensure consistency across health technology appraisals.[32] Literature shows that 5L utility values mapped from the EQ-5D-3L using the copula mapping function of Hernández Alava & Pudney (2017 and 2020) produced substantially different cost-utility estimates compared to 3L value sets.[42–44] As for the crosswalk approach developed by van Hout et al. (2012), one study showed that the use of the crosswalk instead of the England 5L value set may increase the likelihood of mental health interventions being considered cost-effective.[45] It is unclear, however, whether this finding can be applied to other countries that have elicited 5L value sets. In addition, a reverse crosswalk that predicts 3L utility values by mapping EQ-5D-5L to EQ-5D-3L was recently published[41] and information on the impact of using reverse-mapped utility values on cost-utility outcomes is not yet available.

Previous studies also indicate that different country-specific EQ-5D value sets may result in different utility values and QALY estimates.[46–48] However, if the impact is equal in the intervention and control groups, this will not affect incremental QALYs and cost-utility results. Further Investigation of the impact of using different country-specific EQ-5D value sets on cost-utility outcomes is also needed, particularly because some countries do not have value sets available yet and, therefore, use a reference value set from another country.

Predicting health-related quality of life from disease-specific patient-reported outcome measures

In situations that EQ-5D data are not available, utility values might be predicted based on condition-specific patient-reported outcome measures (PROMs). In doing so, two strategies can be used: 1) utility values can be estimated directly using regression modelling techniques (e.g., regression models[49]), and 2) utility values can be estimated indirectly by linking responses on a PROM (i.e., source instrument) to those of the EQ-5D (i.e., target instrument) first and then use this information to estimate utility values (i.e., response mapping approach).[49–52]

One of the most widely used condition-specific PROMs in studies conducted among patients with low back pain (LBP) is the Oswestry Disability Index (ODI).[53,54] Although the EQ-5D and ODI seem to be conceptually linked,[55–57] it is unclear whether the ODI is suitable for predicting missing EQ-5D utility values among patients with chronic LBP when EQ-5D scores are lacking. Previous studies that used ODI scores to predict EQ-5D utility values,[58,59] did not perform a qualitative assessment of the conceptual overlap in the instruments' underlying constructs. Moreover, previous studies used regression modelling techniques, but did not include modelling techniques to account for the ceiling effect of utility values (e.g., Tobit models) or response mapping approaches (e.g., crosswalk, and ordinal logistic regression).[52] Evidence suggests, however, that response mapping approaches perform better than regression modelling techniques. Amongst others, response mapping approaches are thought to be better at preventing regression to the mean,[60] because they aim to align the scales between instruments so that the distributions of their responses are linked.[61,62] Hence, response mapping approaches might result in more valid estimates of individual scores on the target instrument. Up until now, however, it is unclear whether this is also the case for mapping EQ-5D utility values from ODI.

Handling missing data in trial-based economic evaluations

An important methodological challenge in the analysis of trial-based economic evaluations is the handling of missing data. Missing data are common in clinical trials as participants may skip questions, follow-up assessments, and/or drop out of the study.[63] Because costs and QALYs are calculated as the sum of several cost and utility values that are measured at different time points, one missing cost or utility value means that total costs and QALYs cannot be calculated.[11,20,64] This typically leads to high rates of missing data in trial-based economic evaluations.

Historically, missing data in trial-based economic evaluations were handled by simply deleting participants with missing values (i.e., complete-case analysis). However, deleting cases with missing values from the analysis reduces a study's power and potentially biases estimates.[64] Simple imputation methods, such as mean imputation or last observation carried forward, may underestimate the variance in outcome estimates and may lead to bias if dropout is selective (e.g., related to observed information). Advanced imputation methods, such as Multiple Imputation (MI) were found to perform better when handling missing data in trial-based economic evaluations and are therefore increasingly being used.[65–70]

Another strategy to deal with missing data is through the use of Longitudinal Linear Mixed-models (LLM). Although the primary reason to use LMM is to account for multiple measurements within one patient, the maximum likelihood estimation uses all observed data and produces unbiased estimates under the MAR assumption.[71] Particularly in the case of reasonably normally distributed effect outcomes, MI was found to be unnecessary to obtain unbiased estimates.[72,73] Faria et al. (2014) suggested that MI is not required prior to LLM in trial-based economic evaluations either, but this has never been empirically tested.[64] This is important, however, because there are three distinct statistical challenges to trial-based economic evaluations that may affect the performance of LLM when dealing with missing trial-based economic evaluation data: 1) costs are typically heavily right-skewed and QALYs left-skewed; 2) costs and QALYs are cumulative sums over time, and 3) costs and QALYs are correlated. Hence, an investigation of whether MI is necessary prior to LLM in the context of trial-based economic evaluation is needed.

Aims and outline of the thesis

This thesis will address the aforementioned gaps in knowledge, leading to the following methodological research questions:

Health-related quality of life scoring methods:

1a. Does the use of crosswalks instead of EQ-5D value sets impact trial-based economic evaluation results and decision-making? (*Chapters 2* and *3*)

1b. Does the use of different country-specific EQ-5D value sets impact trial-based economic evaluation results and decision-making? (*Chapter 4*)

Predicting health-related quality of life from condition-specific PROMs:

2. Is it valid to predict EQ-5D utility values from a condition-specific PROM for patients with LBP using mapping approaches for use in economic evaluations? (*Chapters 5* and *6*)

Handling missing data in trial-based economic evaluations:

3. Is MI necessary when using a LLM model to estimate cost-utility outcomes and what is the impact on trial-based economic evaluation results and decision-making? (*Chapter 7*)

Addressing these methodological research questions alone, will not improve the conduct of trial-based economic evaluations. To achieve this clear guidelines are needed for researchers on how to optimize the methodological quality of their trial-based economic evaluations. Therefore, *Chapters 8* and *9* include tutorial papers that provide step-by-step guidance for fellow researchers on how to conduct, analyse, and interpret trial-based economic evaluations.

In the general discussion (*Chapter 10*), the main findings of this thesis are discussed in the context of the current health economic literature, and recommendations and implications for further research and practice will be provided.

References

1. Emanuel EJ, Persad G, Upshur R, *et al*. Fair Allocation of Scarce Medical Resources in the Time of Covid-19. *New England Journal of Medicine* 2020; **382**: 2049–55.
2. Alhalaseh YN, Elshabrawy HA, Erashdi M, Shahait M, Abu-Humdan AM, Al-Hussaini M. Allocation of the "Already" Limited Medical Resources Amid the COVID-19 Pandemic, an Iterative Ethical Encounter Including Suggested Solutions From a Real Life Encounter. *Frontiers in Medicine* 2021; **7**. https://www.frontiersin.org/article/10.3389/fmed.2020.616277 (accessed May 12, 2022).
3. Muche-Borowski C, Abiry D, Wagner H-O, *et al*. Protection against the overuse and underuse of health care – methodological considerations for establishing prioritization criteria and recommendations in general practice. *BMC Health Serv Res* 2018; **18**. DOI:10.1186/s12913-018-3569-9.
4. Netherlands S. Dutch health expenditure 10th highest in Europe − CBS. Statistics Netherlands. https://www.cbs.nl/en-gb/news/2020/47/dutch-health-expenditure-10th-highest-in-europe (accessed May 30, 2022).
5. Bakx P, O'Donnell O, van Doorslaer E. Spending on Health Care in the Netherlands: Not Going So Dutch. *Fiscal Studies* 2016; **37**: 593–625.
6. Cahan EM, Kocher B, Bohn R. Why Isn't Innovation Helping Reduce Health Care Costs? | Health Affairs Forefront. 2020; published online June. https://www.healthaffairs.org/do/10.1377/forefront.20200602.168241/full/ (accessed May 30, 2022).
7. Zorguitgaven | Rijksinstituut voor Volksgezondheid en Milieu Ministerie van Volksgezondheid, Welzijn en Sport. https://www.vtv2018.nl.
8. Drummond MF, McGuire A. Economic Evaluation in Health Care: Merging Theory with Practice. Oxford University Press, 2001.
9. Tackling Wasteful Spending on Health | en | OECD. https://www.oecd.org/health/tackling-wasteful-spending-on-health-9789264266414-en.htm (accessed May 12, 2022).
10. Warner MA. Stop Doing Needless Things! Saving Healthcare Resources During COVID-19 and Beyond. *J GEN INTERN MED* 2020; **35**: 2186–8.
11. Drummond MF, Sculpher MJ, Torranc GW. Methods for the economic evaluation of health care programmes, 3rd ed. Oxford: Oxford University Press, 2005.
12. Gray A, Clarke PM, Wolstenholme JL, Wordsworth S. Applied Methods of Cost-effectiveness Analysis in Healthcare. Oxford University Press, 2010 http://econpapers.repec.org/bookchap/oxpobooks/9780199227280.htm (accessed Nov 3, 2016).
13. Charlton V. NICE and Fair? Health Technology Assessment Policy Under the UK's National Institute for Health and Care Excellence, 1999–2018. *Health Care Anal* 2020; **28**: 193–227.
14. Roseboom KJ, van Dongen JM, Tompa E, van Tulder MW, Bosmans JE. Economic evaluations of health technologies in Dutch healthcare decision-making: a qualitative study of the current and potential use, barriers, and facilitators. *BMC Health Serv Res* 2017; **17**. DOI:10.1186/s12913-017-1986-9.
15. Yuba TY, Novaes HMD, de Soárez PC. Challenges to decision-making processes in the national HTA agency in Brazil: operational procedures, evidence use and recommendations. *Health Research Policy and Systems* 2018; **16**: 40.
16. Wang T, McAuslane N, Liberti L, Gardarsdottir H, Goettsch W, Leufkens H. Companies' Health Technology Assessment Strategies and Practices in Australia, Canada, England, France, Germany, Italy and Spain: An Industry Metrics Study. *Frontiers in Pharmacology* 2020; **11**. https://www.frontiersin.org/article/10.3389/fphar.2020.594549 (accessed June 2, 2022).
17. NIHR. National Institute for Health and Care Research. National Institute for Health and Care Research. https://www.nihr.ac.uk/ (accessed June 18, 2022).
18. ZonMw. ZonMw. https://www.zonmw.nl/nl/ (accessed June 18, 2022).

19. Petrou S, Gray A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. *BMJ* 2011; **342**: d1766.

20. Petrou S, Gray A. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. *BMJ* 2011; **342**. DOI:10.1136/bmj.d1548.

21. Gray AM. Cost-effectiveness analyses alongside randomised clinical trials. *Clinical Trials* 2006; **3**: 538–42.

22. Willan AR. Statistical analysis of cost–effectiveness data from randomized clinical trials. *Expert Review of Pharmacoeconomics & Outcomes Research* 2006; **6**: 337–46.

23. Hughes D, Charles J, Dawoud D, *et al.* Conducting Economic Evaluations Alongside Randomised Trials: Current Methodological Issues and Novel Approaches. *PharmacoEconomics* 2016; **34**: 447–61.

24. Campolina AG, Rozman LM, Decimoni TC, Leandro R, Novaes HMD, De Soárez PC. Many Miles to Go: A Systematic Review of the State of Cost-Utility Analyses in Brazil. *Appl Health Econ Health Policy* 2016; published online Oct 31. DOI:10.1007/s40258-016-0290-x.

25. Torrance GW, Thomas WH, Sackett DL. A Utility Maximization Model for Evaluation of Health Care Programs. *Health Serv Res* 1972; **7**: 118–33.

26. Brazier J, Ratcliffe J, Salomon J, Tsuchiya A. Measuring and Valuing Health Benefits for Economic Evaluation, Second Edition. Oxford, New York: Oxford University Press, 2016.

27. Wisløff T, Hagen G, Hamidi V, Movik E, Klemp M, Olsen JA. Estimating QALY Gains in Applied Studies: A Review of Cost-Utility Analyses Published in 2010. *PharmacoEconomics* 2014; **32**: 367–75.

28. Kennedy-Martin M, Slaap B, Herdman M, *et al.* Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. *Eur J Health Econ* 2020; **21**: 1245–57.

29. EuroQol Group. EuroQol-a new facility for the measurement of health-related quality of life. *Health policy* 1990; **16**: 199–208.

30. Herdman M, Gudex C, Lloyd A, *et al.* Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 2011; **20**: 1727–36.

31. Janssen MF, Pickard AS, Golicki D, *et al.* Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Qual Life Res* 2013; **22**: 1717–27.

32. Devlin N, Parkin D, Janssen B. An Introduction to EQ-5D Instruments and Their Applications. In: Devlin N, Parkin D, Janssen B, eds. Methods for Analysing and Reporting EQ-5D Data. Cham: Springer International Publishing, 2020: 1–22.

33. Stolk E, Ludwig K, Rand K, Hout B van, Ramos-Goñi JM. Overview, Update, and Lessons Learned From the International EQ-5D-5L Valuation Work: Version 2 of the EQ-5D-5L Valuation Protocol. *Value in Health* 2019; **22**: 23–30.

34. NWO. Netherlands Code of Conduct for Research Integrity | NWO. 2018. https://www.nwo.nl/en/netherlands-code-conduct-research-integrity (accessed May 23, 2022).

35. Shuster E. Fifty Years Later: The Significance of the Nuremberg Code. *New England Journal of Medicine* 1997; **337**: 1436–40.

36. Resnik DB, Shamoo AE. The Singapore Statement on Research Integrity. *Account Res* 2011; **18**: 71–5.

37. El Alili M, van Dongen JM, Huirne JAF, van Tulder MW, Bosmans JE. Reporting and Analysis of Trial-Based Cost-Effectiveness Evaluations in Obstetrics and Gynaecology. *Pharmacoeconomics* 2017; **35**: 1007–33.

38. Mutubuki EN, El Alili M, Bosmans JE, *et al.* The statistical approach in trial-based economic evaluations matters: get your statistics together! *BMC Health Serv Res* 2021; **21**: 475.

39. van Hout B, Janssen MF, Feng Y-S, *et al.* Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets. *Value in Health* 2012; **15**: 708–15.

40. Hernandez-Alava M, Wailoo A, Pudney S. Methods for mapping between the EQ-5D-5L and the 3L for technology appraisal. 2017; : 35.

41. van Hout BA, Shaw JW. Mapping EQ-5D-3L to EQ-5D-5L. *Value in Health* 2021; **0**. DOI:10.1016/j.jval.2021.03.009.

42. Pennington B, Hernandez-Alava M, Pudney S, Wailoo A. The Impact of Moving from EQ-5D-3L to -5L in NICE Technology Appraisals. *PharmacoEconomics* 2018; published online Aug 9. DOI:10.1007/s40273-018-0701-y.

43. Alava MH, Wailoo A, Grimm S, *et al.* EQ-5D-5L versus EQ-5D-3L: The Impact on Cost Effectiveness in the United Kingdom. *Value in Health* 2018; **21**: 49–56.

44. Wailoo A, Alava MH, Pudney S, *et al.* An International Comparison of EQ-5D-5L and EQ-5D-3L for Use in Cost-Effectiveness Analysis. *Value in Health* 2021; **24**: 568–74.

45. Camacho EM, Shields G, Lovell K, Coventry PA, Morrison AP, Davies LM. A (five-) level playing field for mental health conditions?: exploratory analysis of EQ-5D-5L-derived utility values. *Qual Life Res* 2018; **27**: 717–24.

46. Norman R, Cronin P, Viney R, King M, Street D, Ratcliffe J. International Comparisons in Valuing EQ-5D Health States: A Review and Analysis. *Value in Health* 2009; **12**: 1194–200.

47. Badia X, Roset M, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. *Medical Decision Making* 2001; **21**: 7–16.

48. Gerlinger C, Bamber L, Leverkus F, *et al.* Comparing the EQ-5D-5L utility index based on value sets of different countries: impact on the interpretation of clinical study results. *BMC Research Notes* 2019; **12**: 18.

49. Dakin H, Abel L, Burns R, Yang Y. Review and critical appraisal of studies mapping from quality of life or clinical measures to EQ-5D: an online database and application of the MAPS statement. *Health and Quality of Life Outcomes* 2018; **16**. DOI:10.1186/s12955-018-0857-3.

50. Petrou S, Rivero-Arias O, Dakin H, *et al.* The MAPS Reporting Statement for Studies Mapping onto Generic Preference-Based Outcome Measures: Explanation and Elaboration. *PharmacoEconomics* 2015; **33**: 993–1011.

51. Longworth L, Rowen D. Mapping to Obtain EQ-5D Utility Values for Use in NICE Health Technology Assessments. *Value in Health* 2013; **16**: 202–10.

52. Dakin H. Review of studies mapping from quality of life or clinical measures to EQ-5D: an online database. *Health and Quality of Life Outcomes* 2013; **11**: 151.

53. Froud R, Patterson S, Eldridge S, *et al.* A systematic review and meta-synthesis of the impact of low back pain on people's lives. *BMC Musculoskeletal Disorders* 2014; **15**: 50.

54. Fairbank JCT, Pynsent PB. The Oswestry Disability Index. *Spine* 2000; **25**: 2940–53.

55. Solberg TK, Olsen J-A, Ingebrigtsen T, Hofoss D, Nygaard ØP. Health-related quality of life assessment by the EuroQol-5D can provide cost-utility data in the field of low-back surgery. *Eur Spine J* 2005; **14**: 1000–7.

56. Johnsen LG, Hellum C, Nygaard ØP, *et al.* Comparison of the SF6D, the EQ5D, and the oswestry disability index in patients with chronic low back pain and degenerative disc disease. *BMC Musculoskelet Disord* 2013; **14**: 148.

57. Whynes DK, McCahon RA, Ravenscroft A, Hodgkinson V, Evley R, Hardman JG. Responsiveness of the EQ-5D Health-Related Quality-of-Life Instrument in Assessing Low Back Pain. *Value in Health* 2013; **16**: 124–32.

58. Carreon LY, Bratcher KR, Das N, Nienhuis JB, Glassman SD. Estimating EQ-5D Values From the Oswestry Disability Index and Numeric Rating Scales for Back and Leg Pain. *Spine* 2014; **39**: 678–82.

59. Knio ZO, VanHorn TA, O'Gara TJ. Should the EuroQol-Five Dimensions Replace the Oswestry Disability Index When Tracking Lumbar Tubular Microdecompression Outcomes? *World Neurosurgery* 2020; **134**: e566–71.

60. Thompson NR, Lapin BR, Katzan IL. Mapping PROMIS Global Health Items to EuroQol (EQ-5D) Utility Scores Using Linear and Equipercentile Equating. *Pharmacoeconomics* 2017; **35**: 1167–76.

61. Fayers PM, Hays RD. Should linking replace regression when mapping from profile-based measures to preference-based measures? *Value Health* 2014; **17**: 261–5.

62. Dorans NJ. Linking scores from multiple health outcome instruments. *Qual Life Res* 2007; **16**: 85–94.

63. Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, 2004.

64. Faria R, Gomes M, Epstein D, White IR. A Guide to Handling Missing Data in Cost-Effectiveness Analysis Conducted Within Randomised Controlled Trials. *Pharmacoeconomics* 2014; **32**: 1157.

65. Gabrio A, Mason AJ, Baio G. Handling Missing Data in Within-Trial Cost-Effectiveness Analysis: A Review with Future Recommendations. *Pharmacoecon Open* 2017; **1**: 79–97.

66. Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Sinha D, Parzen M, Lipshultz S. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: An application to AIDS data. *J R Stat Soc Ser A Stat Soc* 2009; **172**: 3–20.

67. Noble SM, Hollingworth W, Tilling K. Missing data in trial-based cost-effectiveness analysis: the current state of play. *Health Economics* 2012; **21**: 187–200.

68. Díaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clin Trials* 2014; **11**: 590–600.

69. Gomes M, Grieve R, Nixon R, Edmunds WJ. Statistical Methods for Cost-Effectiveness Analyses That Use Data from Cluster Randomized Trials: A Systematic Review and Checklist for Critical Appraisal. *Med Decis Making* 2012; **32**: 209–20.

70. Sterne JAC, White IR, Carlin JB, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; **338**: b2393.

71. Gabrio A, Plumpton C, Banerjee S, Leurent B. Linear mixed models to handle missing at random data in trial- based economic evaluations. *Health Economics* 2022; : hec.4510.

72. Peters SAE, Bots ML, Ruijter HM den, *et al.* Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models. *J CLIN EPIDEMIOL* 2012; **65**: 686–95.

73. Twisk J, de Boer M, de Vente W, Heymans M. Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of Clinical Epidemiology* 2013; **66**: 1022–8.

Part 1

Methodological studies

CHAPTER 2

Comparing the EQ-5D-5L crosswalks and value sets for England, the Netherlands, and Spain: exploring their impact on cost-utility results

Ângela Jornada Ben*, Aureliano P. Finch*, Johanna M. van Dongen, Maartje de Wit, Susan E.M. van Dijk, Frank J. Snoek, Marcel C. Adriaanse, Maurits W. van Tulder, Judith E. Bosmans

*contributed equally to this work

Health Economics, 2020; 29(5): 640–651

Abstract

This study compares the EQ-5D-5L crosswalks and the 5L value sets for England, the Netherlands, and Spain and explores the implication of using one or the other for the results of cost-utility analyses. Data from two randomized controlled trials in depression and diabetes were used. Utility value distributions were compared and mean differences in utility values between the EQ-5D-5L crosswalk and the 5L value set were described by country. QALYs were calculated using the area-under-the-curve method. Incremental cost-effectiveness ratios (ICERs) were calculated, and uncertainty around ICERs was estimated using bootstrapping and graphically shown in cost-effectiveness acceptability curves. For all countries investigated, utility value distributions differed between the EQ-5D-5L crosswalk and the 5L value set. In both case studies, mean utility values were lower for the EQ-5D-5L crosswalk compared with the 5L value set in England and Spain, but higher in the Netherlands. However, these differences in utility values did not translate into relevant differences across utility estimation methods in incremental QALYs and the interventions' probability of cost-effectiveness. Thus, our results suggest that EQ-5D-5L crosswalks and 5L value sets can be used interchangeably in patients affected by mild or moderate conditions. Further research is needed to establish whether these findings are generalizable to economic evaluations among severely ill patients.

Introduction

Generic preference-based measures of health (GPBMs) are pre-scored utility measures that can be used to calculate quality-adjusted life-years (QALYs) for economic evaluations.[1] The EQ-5D is the most frequently used GPBM and it is recommended by Health Technology Assessment agencies, such as the National Institute for Health and Care Excellence[2] and the Dutch National Health Care Institute.[3]

The EQ-5D comprises a standardized descriptive system through which health is described, and a value set that reflects the strength of preferences of the general public for the health states described. The original descriptive system of the EQ-5D (commonly referred to as EQ-5D-3L or 3L version) comprised 5 health dimensions (mobility, self-care, usual activities, pain/ discomfort, anxiety/ depression), each of which had three levels of response i.e., no problems, some problems, extreme problems.[4] Although theoretically, the EQ-5D-3L is applicable across all disease areas and conditions, evidence suggests that it lacks responsiveness in specific populations, such as patients with mental health issues, visual disorders, or neoplasms.[1] One possible explanation for this could be the absence of an appropriate number of response levels to capture relevant changes in the patient's health-related quality of life (HRQoL). This has led to the development of a new EQ-5D version; the EQ-5D-5L (commonly referred to as EQ-5D-5L or 5L version). This measure describes health using the same dimensions of the EQ-5D-3L, but it uses five response levels i.e., no problems, slight problems, moderate problems, severe problems, and extreme problems.[5]

The uptake of the EQ-5D-5L amongst researchers has significantly increased in the last years, and numerous economic evaluations in the areas of musculoskeletal diseases, cardiovascular diseases, and mental health are now using this measure.[6–8] Given the importance of employing country-specific value sets because of sociocultural differences among populations,[9] valuation studies of the EQ-5D-5L are being undertaken in many countries. Nonetheless, 5L value sets are not available yet for many countries, while the 3L value sets are. For this reason, many economic evaluations used and plan to use a country-specific interim crosswalk[10] which estimates the EQ-5D-5L utility values from the EQ-5D-3L tariffs.[6,11]

The predictive ability of the crosswalk may be affected by the protocols used for the valuation of the EQ-5D-3L and EQ-5D-5L i.e. the Measurement and Valuation of Health (MVH) for the EQ-5D-3L and the EuroQol Valuation Technology for the EQ-5D-5L.[12,13] Differences in protocols include, for example, the use of different elicitation methods e.g. VAS or DCE, modes of administration e.g. face to face or CAPI and modelling specifications e.g. N3 versus OLS. Moreover, there are also differences in the design of the valuation studies for the same EQ-5D version between countries that might affect the precision of the crosswalk. For example, while in the English and Spanish valuation studies 43 health states were directly valued and 12 health states were administered per responder, the Dutch valuation study directly valued only 17 health states with each responder valuing all of them).[14–16]

Two studies have shown that the EQ-5D-5L crosswalk and 5L value set for England differ in terms of the decrements associated with the response levels and utility range/distribution.[17,18] With the release of 5L value sets for the Netherlands and Spain, it became possible to investigate whether such differences also exist between the crosswalk and value set for other countries. Moreover, the impact of using one utility estimation method over the other on the statistical uncertainty surrounding cost-utility estimates is still unclear. The latter is particularly important as it provides insight into the validity of reimbursement decisions based on economic evaluations that relied on interim crosswalk QALY estimates.[19]

This study builds on the work of Mulhern et al.[17] and Camacho et al.[18] by assessing whether the differences found between the EQ-5D-5L crosswalk and the 5L value set of England, also apply to the Dutch and Spanish crosswalks and value sets. Additionally, this study explores whether the different utility estimation methods have an impact on the outcomes of a cost-utility analysis, such as incremental QALYs and an intervention's probability of being cost-effective. For this purpose, data from two pragmatic randomized clinical trials in patients affected by depression and diabetes were used. These conditions are relevant for the comparison of the crosswalks and the value sets as they are associated with significant impairments in HRQoL.[7,20]

Methods

Data

This study used data from two pragmatic cluster-randomized controlled trials performed in the Netherlands. Full details of the trials are described elsewhere.[8,21] In brief, the first trial (referred to as depression study in this paper) assessed the cost-effectiveness of a program consisting of four sequential treatment steps to prevent major depression in comparison with usual care among patients with subthreshold depression symptoms. In the second trial (referred to as the diabetes study in this paper), the intervention consisted of group sessions aiming at improving symptom recognition and management of hypoglycaemia by patients with diabetes. This intervention was compared to current practice.

Both studies used the Dutch EQ-5D-5L value set for estimating utilities and found that the intervention under study was not cost-effective compared to usual care. In the depression study, the EQ-5D-5L was administered at baseline and every 3 months until a 12-month follow-up. In the diabetes study, it was administered at baseline and at 2, 4, and 6-month follow-ups. Costs were measured from a societal perspective, and included costs of the study interventions, health care utilization, medication, and lost productivity (absenteeism and presenteeism).[8,21]

The EQ-5D-5L utility estimation methods

For both case studies, utility values were estimated using the EQ-5D-5L crosswalk approach[10] and the 5L value sets of England,[22] the Netherlands,[23] and Spain.[24] In the crosswalk method, utility values for the 5L version were predicted from the 3L utilities of the United Kingdom (U.K. crosswalk),[15] the Netherlands (Dutch crosswalk),[16] and Spain (Spanish crosswalk).[14]

Analyses

Hypothetical health states

First, both utility estimation methods were compared in terms of possible changes in utility values between hypothetical adjacent health states. For this purpose, the definition of adjacent health states published by Mulhern et al.[17] was used; i.e., "as having one dimension with one level difference". For instance, in the mobility dimension, the utility differences between health states 11111 and 21111; 21111 and 31111; 31111 and 41111; 41111 and 51111 were calculated. This approach was also applied to the other 4 dimensions, resulting in 25 adjacent hypothetical health states. Subsequently, the utility values of the 25 hypothetical health states were calculated based on both methods and compared. This analysis provides an overview of the potential magnitude of the differences due to the utility estimation method used. We extend the analysis performed by Mulhern et al. by also including the Dutch and Spanish sets.

Case studies – comparison of utility values

Hereafter, the pooled utility value distributions of all measurement points obtained from the samples included in the case studies were compared using Kernel density histograms. Using these data, the mean utility values, standard deviations of the mean utility values (SDs), ranges and 95% confidence intervals (CI) were estimated. Mean differences in utility values between the EQ-5D-5L crosswalk and the 5L value set, including their 95% confidence intervals, were described by country. Also, the utility values generated by both methods were used to calculate QALYs. QALYs were estimated by the area-under-the-curve method with linear interpolation between time points.

Case studies – comparison of cost-utility outcomes

Subsequently, two cost-utility analyses were conducted per country; one using the EQ-5D-5L crosswalk and one using the 5L value set. For both case studies, missing data were imputed using multiple imputation by Chained Equations.[25] The EQ-5D-5L descriptive system was imputed, rather than the corresponding utility values, to allow for the calculation of utility values based on different value sets using the imputed EQ-5D-5L responses. Predictive mean matching was used to deal with the skewed distribution of the costs and the categorical nature of the EQ-5D-5L descriptive system data.[26] All potential variables associated with the "missingness" of data, cost and effect outcomes, and possible confounders were included in the multiple imputation model. Datasets were imputed in order for the loss of efficiency to be less than 5%.[26] Ten datasets were analyzed separately, and estimates were pooled using Rubin's rules.[27]

Incremental costs and incremental QALYs were estimated using seemingly unrelated regression (SUR) analyses,[28] in which two separate regression models for costs and QALYs are estimated and the correlation between costs and QALYs is accounted for through correlated error terms. Bias-corrected and accelerated bootstrapping with 5,000 replications was used to estimate confidence intervals surrounding incremental costs.[29] Incremental Cost-Effectiveness Ratios (ICERs) were calculated by dividing incremental costs by incremental QALYs. Uncertainty surrounding the ICERs was estimated using bootstrapping[30,31] and by plotting cost-effectiveness acceptability curves (CEACs).[32] CEACs indicate the probability of an intervention being cost-effective compared with a control for a range of willingness-to-pay thresholds.[32] All data analyses were performed in Stata Statistical Software 14th version.

Results

Hypothetical health states

Table 1 compares the differences between the EQ-5D-5L crosswalks and the 5L value sets for the 25 adjacent hypothetical health states. The U.K., the Dutch, and the Spanish crosswalks resulted in the largest utility decrements between response levels 5 "extreme" and 4 "severe", whereas the largest decrements in their respective 5L value sets occurred between levels 4 "severe" and 3 "moderate".

In the EQ-5D-5L crosswalk for U.K. the utility decrement between levels 2 "slight" and 1 "no problems" was larger than in the 5L value set for England in all five dimensions. In the Dutch crosswalk, the utility decrement between levels "slight" and "no problems" was very similar and slightly higher as compared to the Dutch 5L value set in all dimensions, except in the usual activities dimension. In the Spanish crosswalk, the utility decrement between levels "slight" and "no problems" was larger (but of small magnitude as compared to the U.K. crosswalk) than in the Spanish 5L value set, except in the anxiety/depression dimension.

Case studies – comparison of utility values

Table S1 presents the background characteristics of the two samples at baseline. Figure 1 shows the kernel density histograms for the EQ-5D-5L crosswalks and the 5L value sets by case study. Both utility estimation methods result in a left-skewed distribution of utility values. In the depression study, the EQ-5D-5L crosswalks and the 5L value sets resulted in a unimodal distribution whereas in the diabetes study they resulted in a bimodal distribution. In both case studies, peaks for the U.K. crosswalk were different than for the England 5L value set. However, there were only slight differences in the utility value distributions for the Dutch and Spanish crosswalks, and their 5L value sets and the peaks for the two methods were similar.

Table 1 | Comparing the utility decrement between adjacent health states by the EQ-5D-5L crosswalk and the 5L value sets of England, the Netherlands, and Spain

	England				The Netherlands				Spain			
Health states	**U.K. Crosswalk**	**Diff**	**5L value set**	**Diff**	**Crosswalk**	**Diff**	**5L value set**	**Diff**	**Crosswalk**	**Diff**	**5L value set**	**Diff**
11111	1.000		1.000		1.000		1.000		1.000		1.000	
21111	0.877	0.123	0.942	0.058	0.912	0.088	0.918	0.082	0.893	0.107	0.916	0.084
31111	0.850	0.027	0.924	0.018	0.893	0.019	0.896	0.022	0.870	0.023	0.901	0.015
41111	0.813	0.037	0.793	**0.131**	0.867	0.026	0.787	**0.109**	0.826	0.044	0.750	**0.150**
51111	0.336	**0.477**	0.726	**0.067**	0.534	**0.333**	0.750	**0.037**	0.255	0.571	0.663	**0.088**
11111	1.000		1.000		1.000		1.000		1.000		1.000	
12111	0.846	0.154	0.950	0.050	0.873	0.127	0.915	0.085	0.868	0.132	0.950	0.050
13111	0.815	0.	0.920	0.030	0.847	0.026	0.892	0.023	0.842	0.026	0.947	0.003
		031										
14111	0.723	0.092	0.836	**0.084**	0.773	0.074	0.785	**0.107**	0.729	0.113	0.836	**0.111**
15111	0.436	**0.287**	0.797	**0.039**	0.543	**0.230**	0.785	**0.000**	0.376	0.353	0.804	**0.032**
11111	1.000		1.000		1.000		1.000		1.000		1.000	
11211	0.906	0.094	0.950	0.050	0.917	0.083	0.914	0.086	0.924	0.076	0.956	0.044
11311	0.883	0.023	0.937	0.013	0.897	0.020	0.866	0.048	0.905	0.019	0.951	0.005
14111	0.776	0.107	0.838	**0.099**	0.812	0.085	0.761	**0.105**	0.769	0.136	0.865	**0.086**
15111	0.556	**0.220**	0.816	**0.022**	0.638	**0.174**	0.761	**0.000**	0.490	0.279	0.847	**0.018**
11111	1.000		1.000		1.000		1.000		1.000		1.000	
11121	0.837	0.163	0.937	0.063	0.874	0.126	0.887	0.113	0.909	0.091	0.922	0.078
11131	0.796	0.041	0.916	0.021	0.843	0.031	0.861	0.026	0.887	0.022	0.899	0.023
11141	0.583	0.213	0.724	**0.192**	0.652	0.191	0.593	**0.268**	0.702	0.185	0.755	**0.144**
11151	0.264	**0.319**	0.665	**0.059**	0.366	**0.286**	0.538	**0.055**	0.424	0.278	0.618	**0.136**
11111	1.000		1.000		1.000		1.000		1.000		1.000	
11112	0.879	0.121	0.922	0.078	0.845	0.155	0.883	0.117	0.932	0.068	0.919	0.081
11113	0.848	0.031	0.896	0.026	0.805	0.040	0.808	0.075	0.914	0.018	0.872	0.047
11114	0.635	0.213	0.715	**0.181**	0.592	0.213	0.597	**0.211**	0.731	0.183	0.730	**0.143**
11115	0.414	**0.221**	0.711	**0.004**	0.370	**0.222**	0.532	**0.065**	0.541	0.190	0.652	**0.077**
55555	-0.594		-0.285		-0.329		-0.446		-0.654		-0.416	

Note. Diff: utility decrement between adjacent health state. *Gray shadow:* In the U.K. crosswalk, the utility decrement between Levels 1 and 2 was larger than in the England 5L value set in all dimensions. In the Dutch crosswalk, the utility decrement between Levels 2 and 1 was very similar and slightly higher as compared with the Dutch 5L value set in all dimensions, except in the usual activities dimension (where the utility decrement is slightly small in the Dutch crosswalk than the Dutch 5L value set). In the Spanish crosswalk the utility decrement between Levels 2 and 1 was larger than in the Spanish 5L value set, except in the anxiety/depression dimension (where the utility decrement is slightly smaller in the Spanish crosswalk than the Spanish 5L value set). *Abbreviations:* 5LEQ-5D-5L, five-level EuroQol five-dimension questionnaire.

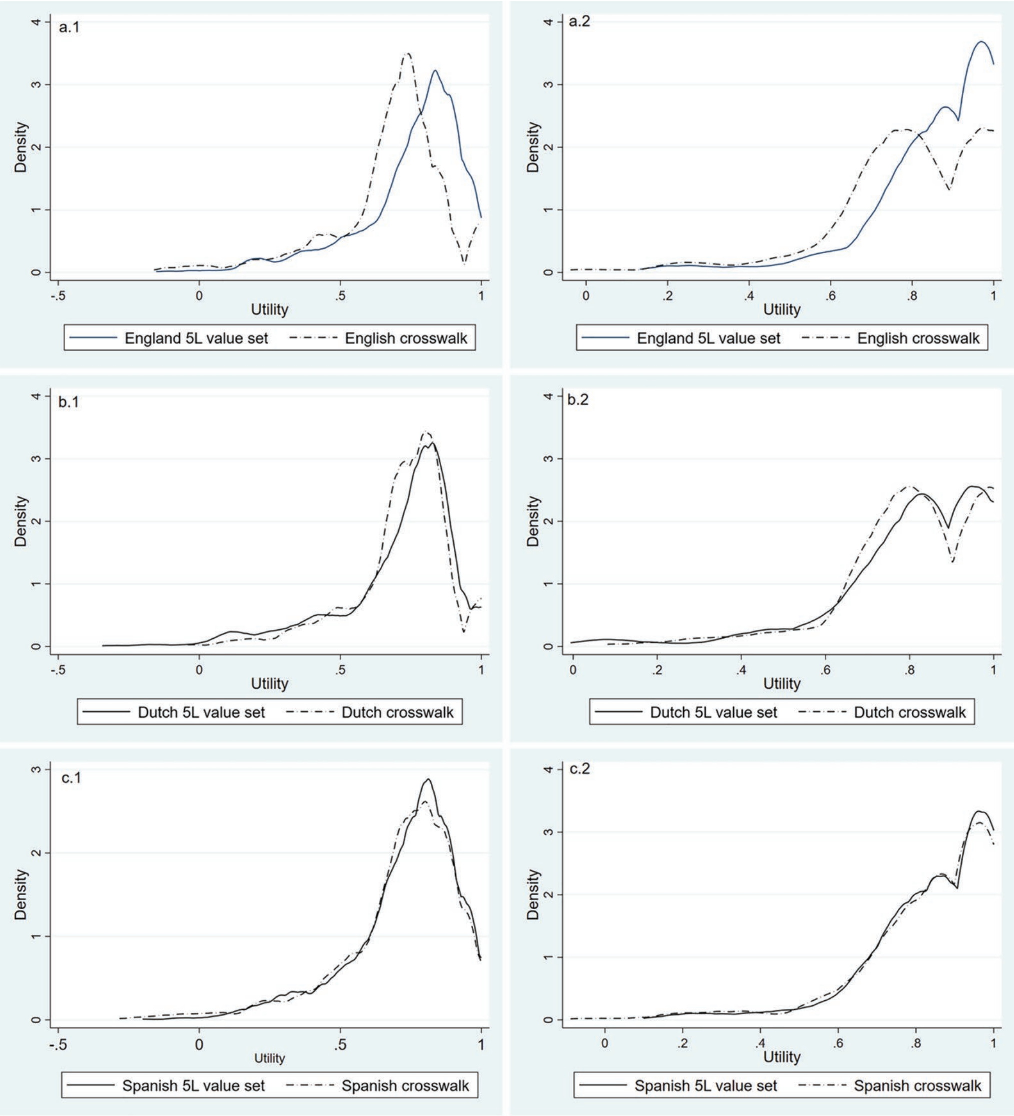


Figure 1 | Kernel density histograms comparing the utility values distribution by the five-level (5L) EuroQol five-dimension questionnaire crosswalks and the 5L value sets by country. Depression study: a.1, b.1, and c.1. Diabetes study: a.2, b.2, and c.2.

Table 2 reports the minimum, maximum, and mean utility values estimated by the EQ-5D-5L crosswalks and the 5L value sets in the two case studies. The maximum utility value was 1.000 for both utility estimation methods in both case studies and for all countries. Differences in minimum utility values were found between the EQ-5D-5L crosswalk and the 5L value set in both case studies, and those differences were most pronounced for the Netherlands. Mean utility values were lower when generated using the EQ-5D-5L crosswalk compared with the 5L value set in both case studies for England and Spain, whereas the opposite was true for the Netherlands.

Case studies – comparison of cost-utility outcomes

Table 3 shows the results of the economic evaluations using both utility estimation methods in each case study per country. In both case studies and for all countries, relatively small differences in incremental QALYs were found between the EQ-5D-5L crosswalk and the 5L value set (≤0.0038). In all scenarios, the intervention was on average less costly and less effective compared to the control condition, resulting in negative ICERs. Thus, both interventions were dominated by the control group for all countries. ICER point estimates differed between the two utility estimation methods for both case studies and all countries, with differences in the magnitude of the ICER point estimates ranging from -35,998€/QALY for Spain to 20,000€/QALY for the Netherlands in the depression study and from -4,063€/QALY for England to 369€/QALY for Spain in the diabetes study. Although the impact of using the EQ-5D-5L crosswalk or the 5L value set on the deterministic estimates of the ICER was large in both case studies and in all countries, this was mainly due to the small differences in QALYs between the treatment groups, and not to the utility estimation methods.

Despite the observed differences in utility value distributions, mean utility values, and ICER point estimates, the distribution of bootstrapped cost-effect pairs on the cost-effectiveness plane was similar for the EQ-5D-5L crosswalks and 5L value sets in both case studies for all countries (Table 3).

Table 4 and Figure 2 show the differences in the probability of cost-effectiveness between both utility estimation methods by country and case study. In both case studies, the probabilities of cost-effectiveness are relatively similar across methods in all three countries for all willingness-to-pay thresholds (Table 4). Consequently, the CEACs of both utility estimation methods look similar for all countries and case studies (Figure 2).

Table 2 | Comparing utility values estimated by the EQ-5D-5L crosswalks and the 5L value sets of England, the Netherlands and Spain

Method	Mean utility (SD)	Min	Max	95% CI
Depression				
U.K. crosswalk	0.676 (0.19)	-0.160	1	0.664; 0.687
England 5L value set	0.752 (0.20)	-0.153	1	0.741; 0.763
Mean difference	**-0.076 (0.02)**			**-0.073; -0.080**
Dutch crosswalk	0.720 (0.17)	-0.041	1	0.710; 0.730
Dutch 5L value set	0.706 (0.22)	-0.344	1	0.694; 0.719
Mean difference	**0.014 (0.07)**			**0.010; 0.018**
Spanish crosswalk	0.712 (0.21)	-0.282	1	0.700; 0.725
Spanish 5L value set	0.730 (0.19)	-0.201	1	0.719; 0.741
Mean difference	**-0.017 (0.07)**			**-0.013; -0.021**
Diabetes study				
U.K. crosswalk	0.808 (0.19)	-0.038	1	0.792; 0.824
England 5L value set	0.862 (0.16)	0.132	1	0.849; 0.875
Mean difference	**-0.054 (0.05)**			**-0.049; -0.059**
Dutch crosswalk	0.8234 (0.17)	0.082	1	0.809; 0.837
Dutch 5L value set	0.8231 (0.19)	-0.006	1	0.808; 0.839
Mean difference	**0.0003 (0.05)**			**0.003; 0.004**
Spanish crosswalk	0.842 (0.18)	-0.088	1	0.827; 0.857
Spanish 5L value set	0.850 (0.16)	0.010	1	0.836; 0.864
Mean difference	**-0.008 (0.05)**			**-0.004; -0.012**

Abbreviations: 5L, five-level; CI, confidence interval; EQ-5D-5L, five-level EuroQol five-dimension questionnaire; *SD*, standard deviation.

Table 3 | Results of the cost-utility analysis

Method	Incremental costs (95% CI)	Incremental QALYs (95% CI)	ICER (€)	Distribution of the Cost-Effectiveness plane (%)			
				North-east[a]	South-east[b]	South-west[c]	North-west[d]
Depression study							
U.K. crosswalk	€2,000 [€-935; €5,122]	-0.0085 [0.0553; 0.0387]	-236,282	28	8	3	61
England 5L value set	€2,000 [€-951; €5,172]	-0.0080 [-0.0526; 0.0365]	-248,856	27	7	3	63
Difference	–	**0.0005**	**-12,574**	**1**	**1**	**0**	**2**
Dutch crosswalk	€2,000 [€-957; €5,159]	-0.0067 [-0.0462; 0.0327]	-297,131	29	7	3	61
Dutch 5L value set	€2,000 [€-970; €5,150]	-0.0072 [-0.058; 0.043]	-277,071	31	7	3	59
Difference	–	**-0.0005**	**20,060**	**2**	**0**	**0**	**2**
Spanish crosswalk	€2,000 [€-935; €5,166]	-0.0084 [-0.0582; 0.0415]	-238,703	28	8	2	61
Spanish 5L value set	€2,000 [€-966; €5,183]	-0.0073 [-0.0530; 0.0385]	-274,700	30	7	3	60
Difference	–	**0.0011**	**-35,997**	**2**	**1**	**1**	**1**
Diabetes study							
U.K. crosswalk	€49 [€-1205; €1,090]	-0.0090 [-0.0378; 0.0197]	-5,441	15	12	33	40
England 5L value set	€49 [€-1206; €1,077]	-0.0052 [-0.0300; 0.0197]	-9,504	19	16	30	35
Difference	–	**0.0038**	**-4,063**	**4**	**4**	**3**	**5**
Dutch crosswalk	€49 [€-1202; €1,097]	-0.0082 [-0.0342; 0.0179]	-6,014	15	12	34	39
Dutch 5L value set	€49 [€-1,205; €1,090]	-0.0066 [-0.363; 0.0231]	-7,424	18	15	31	36
Difference	–	**0.0016**	**-1,410**	**3**	**3**	**3**	**3**
Spanish crosswalk	€49 [€-1204; €1,098]	-0.0071 [-0.0344; 0.0202]	-6,932	16	14	32	38
Spanish 5L value set	€49 [€-1204; €1,098]	-0.0075 [-0.0332; 0.0182]	-6,563	16	13	33	38
Difference	–	**-0.0004**	**369**	**0**	**1**	**1**	**0**

Note: In both case studies, the new intervention was on average more costly and less effective and therefore was dominated by control. *Abbreviations:* CI, confidence interval; ICER, cost-effectiveness ratio; QALY, quality-adjusted life-year.

[a]New intervention is more effective but more costly compared with control. [b]New intervention is more effective and less costly compared with control. [c]New intervention is less effective and less costly compared with control. [d]New intervention is less effective and more costly compared with control.

Table 4 | Comparing the probability of cost-effectiveness between the EQ-5D-5L crosswalk and the 5L value set by country and case study at different willingness-to-pay thresholds

Method	**0 €/QALY gained**	**20,000 €/QALY gained**	**34,000[a] €/QALY gained**
Depression study			
England 5L value set	0.1039	0.1226	0.1382
U.K. crosswalk	0.1041	0.1256	0.1425
Difference	**-0.0002**	**-0.0030**	**-0.0043**
Dutch 5L value set	0.1039	0.1281	0.1480
Dutch crosswalk	0.1025	0.1202	0.1346
Difference	**0.0014**	**0.0079**	**0.0133**
Spanish 5L value set	0.1041	0.1244	0.1413
Spanish crosswalk	0.1032	0.1292	0.1487
Difference	**0.0009**	**-0.0048**	**-0.0073**
Diabetes study			
England 5L value set	0.4668	0.4059	0.3789
U.K. crosswalk	0.4669	0.3644	0.3234
Difference	**-0.0001**	**0.0415**	**0.0554**
Dutch 5L value set	0.4669	0.3915	0.3621
Dutch crosswalk	0.4668	0.3703	0.3284
Difference	**0.0001**	**0.0212**	**0.0337**
Spanish 5L value set	0.4668	0.3783	0.3396
Spanish crosswalk	0.4669	0.3869	0.3536
Difference	**-0.0001**	**-0.0086**	**-0.0140**

Note. The probability of cost-effectiveness is determined as the proportion of the bootstrapped cost-effect pairs where the intervention is cost-effective (i.e., the proportion of the bootstrapped cost-effect pairs falling to the South-East quadrant of the CE plane) given a willingness-to-pay threshold (e.g., the probability of the intervention being cost-effective was.1382 using U.K. crosswalk and.1425 using 5L value set for England at willingness-to-pay of €34,000/QALY gained, a difference of.0043). *Abbreviations:* 5L, five level; EQ-5D-5L, five-level EuroQol five-dimension questionnaire; QALY, quality-adjusted life year.

[a]U.K. upper commonly accepted willingness-to-pay threshold.

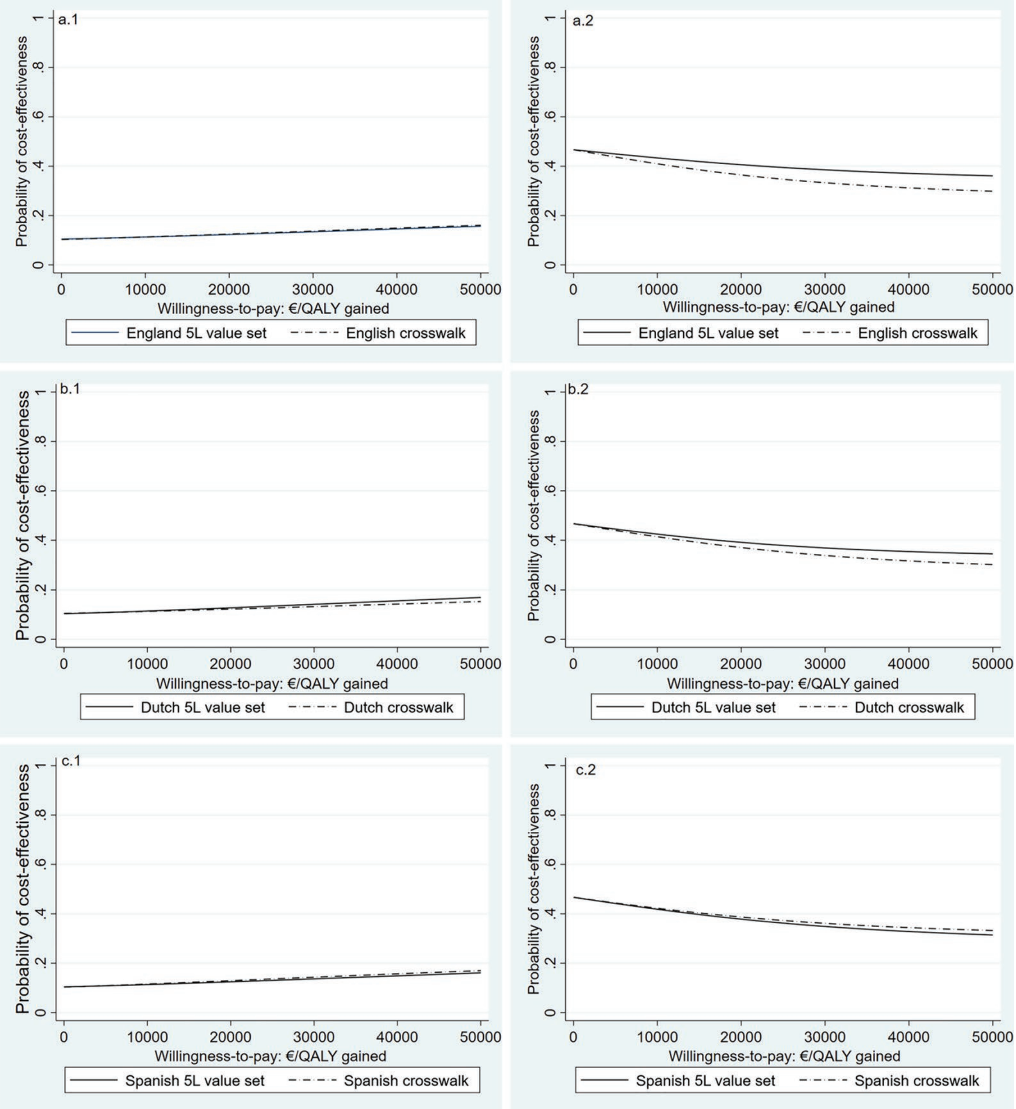


Figure 2 | The five-level (5L) EuroQol five-dimension questionnaire crosswalks and the 5L value sets cost-effectiveness acceptability curves by country and case study. The probability of the intervention being cost-effective at different willingness-to-pay thresholds. QALY, quality-adjusted life year. Depression study: a.1, b.1, and c.1. Diabetes study: a.2, b.2, and c.2

Discussion

The aim of the current study was to compare utility values generated using EQ-5D-5L crosswalks and 5L value sets for England, the Netherlands and Spain, and to assess whether the use of these different utility estimation methods affected cost-utility outcomes in two case studies. Results showed that differences exist in the utility value distribution, mean utility values, and ICER point estimates between the EQ-5D-5L crosswalks and 5L value sets in all countries investigated. However, in our case studies, differences in mean utility values between both utility estimation methods were smaller than the minimum clinically important difference for EQ-5D utility values of 0.074.[33] Moreover, the impact of using either one of those methods on the estimated utility values was similar in the intervention and control groups, thereby not translating into relevant differences in incremental QALYs and probabilities of the interventions being cost-effective.

The observed differences in utility values between the EQ-5D-5L crosswalk and the 5L value set within and between countries could be due to differences in the descriptive system between EQ-5D versions, the use of different modelling techniques, and differences in sociodemographic characteristics of the study populations. It is noteworthy that mean utility values derived using the EQ-5D-5L crosswalk were lower than those derived using the 5L value set in England and Spain, but higher in the Netherlands. As EQ-5D-5L health states and predicted EQ-5D-3L health states were similar in all countries, this is likely due to differences across countries regarding the decrements that the general public attribute to the severity levels of both versions of the EQ-5D. Moreover, although the observed differences did not impact the probability of cost-effectiveness in the uncertainty analyses, deterministic outcomes of economic evaluations were affected by these differences. Especially striking is the large difference in utility values between the U.K. EQ-5D-5L crosswalk and the England 5L value set, which is in contrast with the results for Spain and the Netherlands. This finding supports the recommendation of NICE to calculate utility values using the crosswalk from the 3L value set and not the EQ-5D-5L values.[34]

To the best of our knowledge, only three studies have been published comparing the EQ-5D-5L crosswalk and 5L value set.[17,18,35] The first study, conducted by Mulhern et al., (2018) observed that the England 5L value set produces utility values that are on average 0.085 points higher than the U.K. EQ-5D-5L crosswalk across 12 health conditions, including depression and diabetes. In the current study, we observed similar differences in utility values for England (e.g., a 0.076 points difference among depressed patients), but also that these differences were less pronounced for the Netherlands and Spain.

The second study, conducted by Camacho et al., found mean utility values estimated using the England 5L value set to be approximately 0.08 points higher than mean utility values estimated using the U.K. EQ-5D-5L crosswalk in people with mental health issues. Consequently, they found ICER point estimates to be higher for the U.K. EQ-5D-5L crosswalk than for the England 5L value set.[18] In our study, similar differences in mean utility values and ICER point estimates were found across utility estimation methods in both case studies for all countries investigated.

More recently, Yang et al.[35] explored the impact of using EQ-5D-5L crosswalks or 5L value sets on incremental QALYs in a model-based economic evaluation for seven countries (including England and the Netherlands, but not Spain). For this, they used data from patients at end-stage renal disease. Similar to our results, they found that both utility estimation methods resulted in comparable incremental QALYs for the Netherlands (mean difference: 0.009). However, they found a relatively large difference in incremental QALYs between the U.K. EQ-5D-5L crosswalk and England 5L value set (mean difference: 0.098), which was in contrast with our results. This difference in results may be due to the fact that Yang et al. included severely ill patients, whereas moderately ill patients were included in our case studies.

The current study went beyond the aforementioned studies by assessing the impact of using either one of the utility estimation methods on utility values as well as cost-utility analysis outcomes, such as incremental QALYs, ICER point estimates, and CEACs. Our results indicated that the identified differences in the utility value distribution, mean utility values, and ICER point estimates between the EQ-5D-5L crosswalks and the 5L value sets for England, the Netherlands, and Spain did not translate into relevant differences in incremental QALYs and the probability of the interventions being cost-effective as compared to the control group. This is likely caused by the high level of uncertainty surrounding utility values and QALYs, and more importantly, by the fact that the impact of using either one of the utility estimation methods was similar in the intervention and control groups, thereby not affecting cost-utility outcomes, such as incremental QALYs and CEACs. However, it is unclear whether this also applies to populations with more severe health conditions.

This study has some limitations. First, both case studies included relatively few patients with more severe EQ-5D-5L health states, and the interventions under study were primarily aimed at improving quality of life, instead of increasing life expectancy. Therefore, further research is needed to assess whether the current findings also apply to interventions aimed at more severely ill patients. This is particularly important because utility decrements between response levels 5 and 4 are typically larger in the EQ-5D-5L crosswalks than in the 5L value sets. Another potential limitation is that the difference in QALYs between the intervention and control groups was relatively small and in favour of the control group in both case studies. Future research should indicate whether the probability of an intervention being cost-effective differs in economic evaluations where the difference in QALYs across treatment groups is larger and in favour of the intervention group.

The strengths of our study are that we did not only explore differences in utility values but also performed a full cost-utility analysis to assess the impact on cost-utility outcomes, i.e., incremental QALYs and ICERs. Moreover, to the best of our knowledge, we are the first to assess the joint uncertainty around ICER in two empirical longitudinal datasets in relation to the use of these different utility estimation methods. Bearing in mind the importance of using country-specific value sets in order to account for sociocultural differences among populations,[9] another strength is that we explored the impact of using the crosswalk or value set in three different countries.

Conclusions

In line with previous research, the current study found differences to exist in utility value distributions, mean utility values, and ICER point estimates between EQ-5D-5L crosswalks and 5L value sets for England, the Netherlands, and Spain. However, in our case studies, these differences did not translate into relevant differences in incremental QALYs and the probability of the intervention being cost-effective as compared to the control group at different willingness-to-pay thresholds. In spite of the identified differences between both utility valuation methods, this suggests that EQ-5D-5L crosswalks and 5L value sets can be used interchangeably in economic evaluations of patients affected with mild or moderate conditions. Further research is needed to establish whether these findings also apply in situations where the EQ-5D-5L is administered to severely ill patients.

References

1. Finch AP, Brazier JE, Mukuria C. What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. *Eur J Health Econ* 2017; **19**: 1–14.
2. NICE. Position statement on use of the EQ-5D-5L value set for England (updated October 2019) | NICE technology appraisal guidance | NICE guidance | Our programmes | What we do | About | NICE. 2019 https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l.
3. ZIN. Zorginstituut Nederland − Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg. Ministerie van Volksgezondheid, Welzijn en Sport. 2016; published online Feb 29. https://www.zorginstituutnederland.nl/publicaties/publicatie/2016/02/29/richtlijn-voor-het-uitvoeren-van-economische-evaluaties-in-de-gezondheidszorg (accessed March 28, 2018).
4. EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy* 1990; **16**: 199–208.
5. Herdman M, Gudex C, Lloyd A, *et al.* Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 2011; **20**: 1727–36.
6. Jensen CE, Riis A, Pedersen KM, Jensen MB, Petersen KD. Study protocol of an economic evaluation of an extended implementation strategy for the treatment of low back pain in general practice: a cluster randomised controlled trial. *Implementation Science* 2014; **9**: 2–7.
7. Ruo B, Rumsfeld JS, Hlatky MA, Liu H, Browner WS, Whooley MA. Depressive Symptoms and Health-Related Quality of Life: The Heart and Soul Study. *JAMA* 2003; **290**: 215–21.
8. van Dijk SE, Pols AD, Adriaanse MC, *et al.* Cost-effectiveness of a stepped-care intervention to prevent major depression in patients with type 2 diabetes mellitus and/or coronary heart disease and subthreshold depression: design of a cluster-randomized controlled trial. *BMC Psychiatry* 2013; **13**: 128.
9. Xie F, Gaebel K, Perampaladas K, Doble BM, Pullenayegum E. Comparing EQ-5D valuation studies: A systematic review and methodological reporting checklist. *Value in Health* 2013; **16**: A44–5.
10. van Hout B, Janssen MF, Feng Y-S, *et al.* Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets. *Value in Health* 2012; **15**: 708–15.
11. Krog AH, Sahba M, Pettersen EM, Wisløff T, Sundhagen JO, Kazmi SS. Cost-utility analysis comparing laparoscopic vs open aortobifemoral bypass surgery. *Vasc Health Risk Manag* 2017; **13**: 217–24.
12. Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A Program of Methodological Research to Arrive at the New International EQ-5D-5L Valuation Protocol. *Value in Health* 2014; **17**: 445–53.
13. Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goñi JM, Luo N. EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. *PharmacoEconomics* 2016; **34**: 993–1004.
14. Badia X, Roset M, Herdman M, Kind P. A Comparison of United Kingdom and Spanish General Population Time Trade-off Values for EQ-5D Health States. *Medical decision making* 2001; **21**: 7–16.
15. Dolan PDp. Modeling Valuations for EuroQol Health States : Medical Care. *Medical Care* 1997; **35**: 1095–108.
16. Lamers LM, McDonnell J, Stalmeier PFM, Krabbe PFM, Busschbach JJV. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. *Health Economics* 2006; **15**: 1121–32.
17. Mulhern B, Feng Y, Shah K, *et al.* Comparing the UK EQ-5D-3L and English EQ-5D-5L Value Sets. *PharmacoEconomics* 2018; **36**: 699–713.
18. Camacho EM, Shields G, Lovell K, Coventry PA, Morrison AP, Davies LM. A (five-) level playing field for mental health conditions?: exploratory analysis of EQ-5D-5L-derived utility values. *Qual Life Res* 2018; **27**: 717–24.

19. Brazier J, Briggs A, Bryan S. EQ-5D-5L: Smaller steps but a major step change? *Health Economics* 2018; **27**: 4–6.

20. Evans Kreider K, Pereira K, Padilla BI. Practical Approaches to Diagnosing, Treating and Preventing Hypoglycemia in Diabetes. *Diabetes Ther* 2017; **8**: 1427–35.

21. Wit M, Rondags SMPA, Tulder MW, Snoek FJ, Bosmans JE. Cost-effectiveness of the psycho-educational blended (group and online) intervention HypoAware compared with usual care for people with Type 1 and insulin-treated Type 2 diabetes with problematic hypoglycaemia: analyses of a cluster-randomized controlled trial. *Diabetic Medicine* 2018; **35**: 214–22.

22. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Economics* 2017; : 1–16.

23. M. Versteegh M, M. Vermeulen K, M. A. A. Evers S, de Wit GA, Prenger R, A. Stolk E. Dutch Tariff for the Five-Level Version of EQ-5D. *Value in Health* 2016; **19**: 343–52.

24. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. *Medical Care* 2017; **55**: e51.

25. van Buuren S. Flexible Imputation of Missing Data. Taylor & Francis Group. Chapman & Hall/CRC, 2018 https://stefvanbuuren.name/fimd/ (accessed July 14, 2020).

26. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 2011; **30**: 377–99.

27. Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, 2004.

28. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics* 2004; **13**: 461–75.

29. Chaudhary M, Stearns S. Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Statist Med* 1996; **15**: 1447–58.

30. Black WC. The CE plane: a graphic representation of cost-effectiveness. *Med Decis Making* 1990; **10**: 212–4.

31. Barber JA, Thompson SG. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine* 2000; **19**: 3219–36.

32. Fenwick E, O'Brien BJ, Briggs A. Cost-effectiveness acceptability curves – facts, fallacies and frequently asked questions. *Health Econ* 2004; **13**: 405–15.

33. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Qual Life Res* 2005; **14**: 1523–32.

34. NICE. Position statement on use of the EQ-5D-5L value set for England (updated October 2019) | NICE technology appraisal guidance. NICE, 2019 https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l.

35. Yang F, Devlin N, Luo N. Cost-Utility Analysis Using EQ-5D-5L Data: Does How the Utilities Are Derived Matter? *Value Health* 2019; **22**: 45–9.

Supplementary Table 1 | Characteristics of samples from both case studies

	Depression study n=236	Diabetes study n=137
Age, mean (SD)	**67 (10.0)**	**52 (13.2)**
Female, n (%)	**107 (45.3)**	**63 (46.0)**
Mobility		
None	76 (35.0)	100 (73.0)
Slight	62 (28.5)	27 (19.7)
Moderate	47 (21.6)	4 (2.9)
Severe	30 (13.8)	5 (3.6)
Extreme/ unable to	2 (0.9)	1 (0.7)
Self-care		
None	188 (86.6)	128 (93.4)
Slight	18 (8.3)	8 (5.8)
Moderate	8 (3.7)	1 (0.7)
Severe	1 (0.5)	0
Extreme/ unable to	2 (0.9)	0
Usual activities		
None	68 (31.3)	79 (57.7)
Slight	93 (42.9)	24 (17.5)
Moderate	41 (18.9)	28 (20.4)
Severe	15 (6.9)	4 (2.9)
Extreme/ unable to	0	2 (1.5)
Pain/ discomfort		
None	32 (14.7)	73 (53.3)
Slight	89 (41.0)	35 (25.5)
Moderate	71 (32.7)	22 (16.1)
Severe	22 (10.1)	6 (4.4)
Extreme/ unable to	3 (1.4)	1 (0.7)
Anxiety/ depression		
None	105 (48.4)	74 (54.0)
Slight	79 (36.4)	39 (28.5)
Moderate	29 (13.4)	17 (12.4)
Severe	4 (1.8)	7 (5.1)
Extreme/ unable to	0	0

CHAPTER 3

To what extent does the use of crosswalks instead of EQ-5D value sets impact reimbursement decisions? A simulation study

Ângela Jornada Ben, Aureliano P. Finch, Johanna M. van Dongen,
Mohamed El Alili, Judith E. Bosmans

European Journal of Health Economics, 2022

Abstract

Purpose

Inconsistent results have been found on the impact of using crosswalks versus EQ-5D value sets on reimbursement decisions. We sought to further investigate this issue in a simulation study.

Methods

Trial-based economic evaluation data were simulated for different conditions (depression, low back pain, osteoarthritis, cancer), severity levels (mild, moderate, severe), and effect sizes (small, medium, large). For all 36 scenarios, utilities were calculated using 3L and 5L value sets and crosswalks (3L to 5L and 5L to 3L crosswalks) for the Netherlands, the United States, and Japan. Utilities, quality-adjusted life years (QALYs), incremental QALYs, incremental cost-effectiveness ratios (ICERs), and probabilities of cost-effectiveness (pCE) obtained from values sets and crosswalks were compared.

Results

Differences between value sets and crosswalks ranged from -0.33 to 0.13 for utilities, from -0.18 to 0.13 for QALYs, and from -0.01 to 0.08 for incremental QALYs, resulting in different ICERs. For small effect sizes, at a willingness-to-pay of €20,000/QALY, the largest pCE difference was found for moderate cancer between the Japanese 5L value set and 5L to 3L crosswalk (difference=0.63). For medium effect sizes, the largest difference was found for mild cancer between the Japanese 3L value set and 3L to 5L crosswalk (difference=0.06). For large effect sizes, the largest difference was found for mild osteoarthritis between the Japanese 3L value set and 3L to 5L crosswalk (difference=0.08).

Conclusion

The use of crosswalks instead of EQ-5D value sets can impact cost-utility outcomes to such an extent that this may influence reimbursement decisions.

Introduction

The EQ-5D is one of the most frequently used generic preference-based measures of health-related quality of life in economic evaluations worldwide,[1,2] as it is shown to be valid and responsive in multiple health conditions[3,4] and cultural contexts.[5] It comprises a standardized descriptive system that describes health using five health dimensions (i.e., mobility, self-care, usual activities, pain/discomfort, and anxiety/depression). The original EQ-5D uses 3 severity levels per health dimension (EQ-5D-3L) to describe an individual's health state, that is "no problems", "some problems", and "extreme problems" (further referred to as the EQ-5D-3L).[6] To increase its sensitivity to changes within and between subjects' health states and to reduce commonly observed ceiling effects, a 5-level version of the EQ-5D was developed (further referred to as the EQ-5D-5L).[7,8] The EQ-5D-5L describes health in terms of the same health dimensions, but uses 5 severity levels, that is "no problem", "slight problems", "moderate problems", "severe problems", and "extreme problems". Literature has shown that the EQ-5D-5L has improved measurement properties compared with the EQ-5D-3L.[9–11]

For Health Technology Assessment (HTA) purposes, EQ-5D health states are preferably scored using country-specific value sets. A value set includes a number of utilities assigned to each of the health states described by the EQ-5D.[12] These utilities typically indicate the general public's preferences for a certain health state on a scale anchored at 0 (equaling death) and 1.0 (equaling full health). Utilities below zero are possible for health states that are considered to be worse than dead. By multiplying these utilities by the duration an individual spends in a certain health state, quality-adjusted life years (QALY) are calculated, which is the main effect outcome in cost-utility analyses.[13]

In many countries, value sets are available for the EQ-5D-3L and/or the EQ-5D-5L. The use of national EQ-5D value sets is advised, if they have been produced according to the latest standard procedures (e.g., the EuroQol Valuation Technology – EQ-VT – protocol).[14,15] Otherwise, the country-specific value set may not be recommended to be used by HTA agencies. For example, the National Institute for Health and Care Excellence (NICE) currently does not recommend using the EQ-5D-5L value set for England[15] due to methodological issues found in the initial version of the EQ-VT protocol,[15,17] but to use the mapping approach developed by Hernández-Alava & Pudney (2017) as an interim scoring method instead.[18,19] In other situations, EQ-5D-3L or EQ-5D-5L data may have been collected in a clinical trial, while there is no national value set available at all for the country in which the trial was performed. In those cases, researchers may use a reference value set close to the socio-cultural context of application. It may also happen that a value set is only available for one of the EQ-5D versions (e.g., 3L), while data have been collected using the other version (e.g., 5L). In most of these cases, mapping approaches, such as crosswalks and copula mapping models, can be used to estimate utilities for the other instrument.[20–22] The most widely used mapping approach for HTA purposes[23] is the one of van Hout et al. (2012),[20] which estimates 5L utilities by mapping EQ-5D-5L to EQ-5D-3L (i.e., 5L to 3L crosswalk). An extension of this mapping approach was recently published by van Hout and Shaw (2021),[22] which estimates 3L utilities from mapping EQ-5D-3L to EQ-5D-5L (i.e., 3L to 5L crosswalk).

Given that healthcare decision-makers can be confronted with scientific evidence that is based on EQ-5D value sets or mapping approaches, guidance on choosing the most appropriate utility scoring method is urgently needed.[23] So far, the literature suggests that EQ-5D scoring methods might result in different utility values, but inconsistent results have been found on the extent to which these differences affect differences in QALY between treatment groups (i.e., incremental QALY) and impact reimbursement decisions.[18,24–29] Camacho et al. (2018), for example, concluded that the use of crosswalks instead of England 5L value sets may increase the likelihood of mental health interventions being cost-effective, while Ben et al. (2020) found that the probability of interventions for mental health and diabetes being cost-effective was not significantly affected by using crosswalks compared to 5L value sets for England, the Netherlands, and Spain. Both studies, however, only used data of a small number of empirical studies (i.e., ≤5), which typically assessed a restricted number of health conditions and interventions with relatively small effect sizes.

This study was, therefore, conducted to further investigate the impact of using the 5L to 3L crosswalk compared to 5L value sets on cost-utility outcomes, and hence the possible impact on reimbursement decisions, in a broad range of simulated scenarios. These scenarios included a broader range of health conditions, particularly those that are associated with moderate and severe EQ-5D health states. Moreover, as a 3L to 5L crosswalk[22] has recently been published, we also decided to assess the impact of using the 3L to 5L crosswalk compared to the 3L value set in a wide range of simulated scenarios.

Methods

To evaluate the impact of using crosswalks or EQ-5D value sets on cost-utility outcomes, trial-based economic evaluation data were simulated. In total, 36 different scenarios were simulated including four health conditions (i.e., depression, low back pain, osteoarthritis, and cancer), three severity levels (i.e., mild, moderate, and severe), and three treatment effect sizes (i.e., small, medium, and large). An overview of all scenarios can be found in Table 1. After using four EQ-5D scoring methods to estimate utilities (i.e., 3L and 5L value sets, 3L to 5L and 5L to 3L crosswalks) for the Netherlands (NL), the United States (US), and Japan (JP), cost-utility analyses were performed for all 36 scenarios. Finally, results obtained from the country-specific EQ-5D value sets and mapping approaches (also referred to as 3L to 5L and 5L to 3L crosswalks in this paper) were compared.

Data generation

Data from eight trial-based economic evaluations were used to inform the data generation process. These datasets contained EQ-5D-3L and EQ-5D-5L data of patients with depression,[29,30] low back pain,[31,32] osteoarthritis,[33,34] and cancer.[35,36]

First, the probabilities of observing the different EQ-5D-3L and EQ-5D-5L response levels per health dimension at baseline were extracted from the empirical data by treatment group (i.e., intervention and control). This was done for each EQ-5D version, health condition, and severity level separately. An overview of the cut-off scores[30,37–44] used to classify patients as either having mild, moderate, or severe symptoms per health condition can be found in Appendix 1. Based on the extracted baseline probabilities, 150 baseline profiles were generated for a hypothetical intervention and control group. This was done using the EQ-5D simulation laboratory R package developed by Parkin et al., which is provided the EuroQol Foundation for simulation studies.[45] This package allows researchers to generate datasets with EQ-5D health states (e.g., 12312) of artificial patients, based on pre-specified probabilities of observing the specific response levels within the dimensions. In the current study, these probabilities were based on empirical datasets.[29–36]

Subsequently, 150 follow-up profiles were generated by treatment group for each EQ-5D version, health condition, and severity level separately. This was done using a matrix of transition probabilities which were also based on the empirical datasets.[29–36] These transitions probabilities were then tweaked to obtain small, medium, and large treatment effect sizes. The magnitude of the effect sizes was based on Cohen's d (0.1-0.3 small, 0.5-0.7 medium, and >0.8 large).[46]

Finally, baseline characteristics (i.e., age and gender) and follow-up costs were generated and linked to the health profiles using the simstudy R package.[47] Age was generated from a uniform integer distribution including minimum and maximum values of 25 and 75 years, respectively. The proportion of male subjects was randomly generated from a binary distribution with a mean of 0.19. Follow-up costs were generated from a gamma distribution with a mean of €2000, a "true value" of the mean difference between treatment groups of €250, and a variance of 1. Please note that "true value" means that in 95% of the cases, €250 is included in the 95% confidence interval of the generated cost difference. A negative correlation between costs and QALYs was implemented (r≈-0.10). This means that high costs are associated with lower QALYs and vice-versa. The R script for the data generation can be found at https://github.com/angelajben/EQ5D-simulation-crosswalks or in Appendix 2.

Scoring methods

Utilities were estimated using four EQ-5D scoring methods: 3L value set, 5L value set, 3L to 5L crosswalk,[22] and 5L to 3L crosswalk.[20] For both versions of the EQ-5D, utilities were calculated for NL, US, and JP using the eq5d R package.[48] These three countries were chosen, because they differ considerably in terms of the utility decrements assigned to the different health dimensions of the EQ-5D. For example, for the EQ-5D-3L, the decrement of being "confined to bed" (response level 3 on the mobility dimension) is 0.161 in NL, 0.490 in US and 0.418 in JP. Another example is the decrement of being "extremely anxious or depressed" of the EQ-5D-5L (response level 5 in the anxiety/depression dimension), which is 0.421 in NL, 0.340 in US, and 0.197 in JP. Subsequently, 3L to 5L and 5L to 3L crosswalked utilities for the three countries were estimated using the mapping approaches available on the EuroQol website: https://euroqol.org/support/analysis-tools/cross-walk/. These mapping approaches were chosen as they are the ones mostly used in practice.[23]

Analysis

Utilities and QALYs

For all scenarios and countries, the utilities distribution of the two simulated measurement points (i.e., baseline and follow-up) were assessed using Kernel density histograms. Additionally, mean utilities at baseline and mean QALYs (estimated using the area under the curve method)[13] as well as their respective standard deviations and ranges were described. For the EQ-5D-3L, utilities and

QALYs estimated using country-specific 3L value sets and their respective 3L to 5L crosswalks were described. For the EQ-5D-5L, utilities and QALYs estimated using the country-specific 5L value sets and their respective 5L to 3L crosswalks were described. Differences in utilities and QALYs between EQ-5D value sets and mapping approaches were compared using paired t-tests and their corresponding 95% confidence intervals (95% CI) were described per country. To explore whether the differences between scoring methods were clinically relevant, a minimally clinically important difference of 0.074 was used as a threshold.[49]

Cost-utility analysis

Using QALYs derived from the four EQ-5D scoring methods, cost-utility analyses were performed for all 36 scenarios per country. Incremental QALYs and costs between treatment groups and surrounding 95% CIs were estimated using seemingly unrelated regression analyses.[50] Incremental cost-effectiveness ratios (ICERs) were calculated by dividing incremental costs by incremental QALYs. Bias-corrected and accelerated bootstrapping with 2,000 replications was used to estimate statistical uncertainty surrounding the ICERs.[51,52] The distribution of the bootstrapped estimates was presented in the cost-effectiveness plane (CE-plane).[51] The probability of an intervention being cost-effective compared to control was estimated using the Incremental Net Benefit (INB) approach, where the probability of cost-effectiveness was estimated as the probability that INB>0 for every value of the willingness-to-pay (WTP) threshold (i.e., €0, €20,000, €30,000, and €50,000 per QALY).[53] In this study, an intervention was considered cost-effective if the probability of cost-effectiveness at a specific WTP threshold was ≥0.80. Cost-utility analysis outcomes were descriptively compared across scoring methods (i.e., between EQ-5D value sets and crosswalks). Data analyses were performed in StataSE 16® (StataCorp LP, CollegeStation, TX, US).

Table 1 | Overview of simulated scenarios

Scenario	**Patient population**		Effect size
	Health condition	**Severity level**	
(1)	**Depression**	Mild	Small
(2)			Medium
(3)			Large
(4)		Moderate	Small
(5)			Medium
(6)			Large
(7)		Severe	Small
(8)			Medium
(9)			Large
(10)	**Low back pain**	Mild	Small
(11)			Medium
(12)			Large
(13)		Moderate	Small
(14)			Medium
(15)			Large

Scenario	Patient population		Effect size
	Health condition	**Severity level**	
(16)		Severe	Small
(17)			Medium
(18)			Large
(19)	**Osteoarthritis**	Mild	Small
(20)			Medium
(21)			Large
(22)		Moderate	Small
(23)			Medium
(24)			Large
(25)		Severe	Small
(26)			Medium
(27)			Large
(28)	**Cancer**	Mild	Small
(29)			Medium
(30)			Large
(31)		Moderate	Small
(32)			Medium
(33)			Large
(34)		Severe	Small
(35)			Medium
(36)			Large

Third-six different scenarios were simulated including four different conditions (i.e., depression, low back pain, osteoarthritis, and cancer), three severity levels (i.e., mild, moderate, and severe health states), and three treatment effect sizes (i.e., small, medium, and large) for the Netherlands, the United States and Japan.

Results

Utilities

The distribution of utilities at baseline estimated by the crosswalks differed in all scenarios and countries from those estimated by 3L and 5L value sets. Differences in utilities distributions were more pronounced for the EQ-5D-3L than for the EQ-5D-5L. An example of such differences is shown in Figures 1 and 2. Detailed information can be found in Appendix 3.

Differences in baseline utilities between EQ-5D value sets and crosswalks ranged from -0.33 for the severe low back pain scenario (i.e., between the US 5L value set and 5L to 3L crosswalk, Table 3) to 0.13 for severe cancer scenario (i.e., between the US 3L value set and 3L to 5L crosswalk, Table 2). Baseline utilities estimated by EQ-5D value sets differed statistically significantly from those estimated using crosswalks in all health conditions and severity levels in the investigated countries, except for the Dutch EQ-5D-3L estimates for severe osteoarthritis (difference=0.001, IC 95% -0.01; 0.01, Table 2) and for the Japanese EQ-5D-5L estimates for moderate depression (difference=-0.002, IC 95% -0.01; 0.003, Table 3).

No clinically relevant differences between the Japanese 3L value set and 3L to 5L crosswalk were found, whereas clinically relevant differences were found in 17% of the 12 possible comparisons between the Dutch 3L value set and 3L to 5L crosswalk and in 67% of those between the US value set and 3L to 5L crosswalk (Table 2). No clinically relevant differences between the Japanese 5L value set and 5L to 3L crosswalk were found, whereas, between the Dutch and US value sets and their respective 5L to 3L crosswalks, clinically relevant differences were found in 33% and 50% of the comparisons, respectively (Table 3).

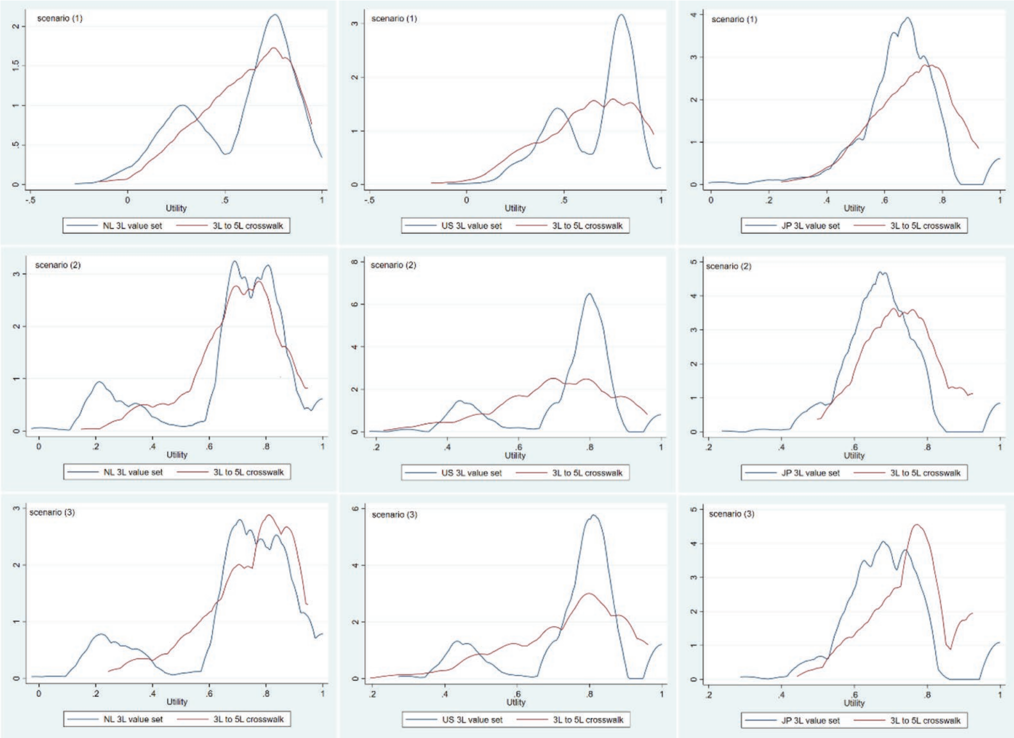


Figure 1 | Utility distribution EQ-5D-3L value sets and 3L to 5L crosswalks for the Netherlands (NL), the United States (US), and Japan (JP). Scenario 1): mild depression and small treatment effect size. Scenario 2): mild depression and medium treatment effect size. Scenario 3): mild depression and large treatment effect size.

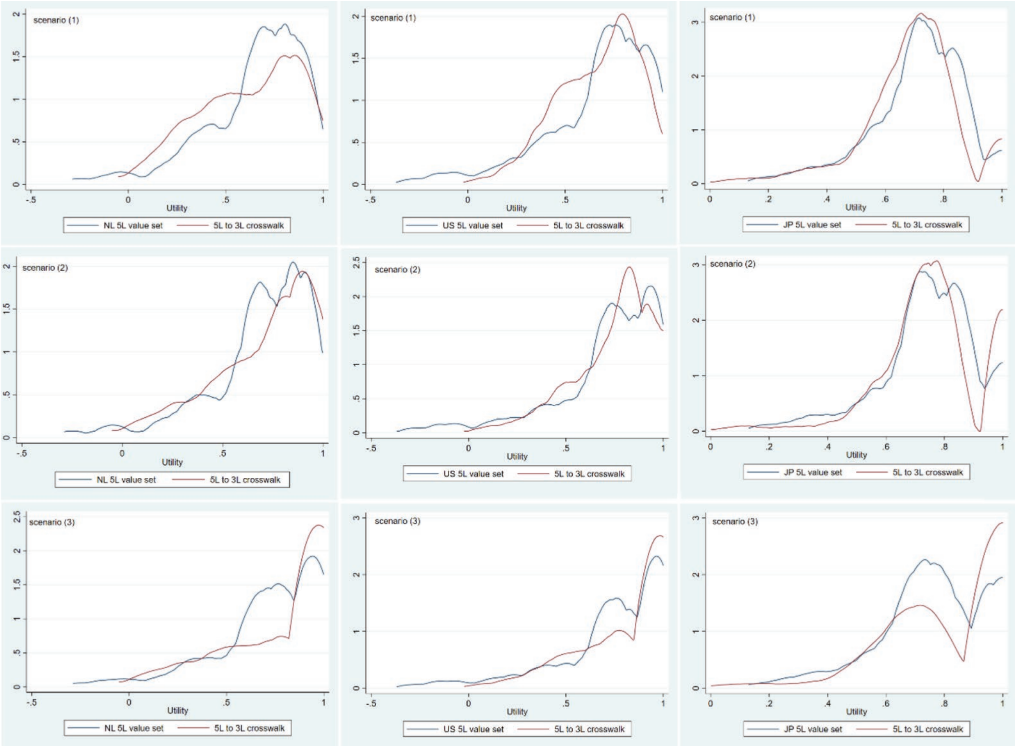


Figure 2 | Utility distribution EQ-5D-5L value sets and 5L to 3L crosswalks for the Netherlands (NL), the United States (US), and Japan (JP). Scenario 1): mild depression and small treatment effect size. Scenario 2): mild depression and medium treatment effect size. Scenario 3): mild depression and large treatment effect size.

Table 2 | Differences in utilities estimated by 3L value sets and 3L to 5L crosswalks

Country	Scoring method	Patient population	Mean utilities (SD)	Min	Max	3Lvs – 3L to 5L cw (95% CI)
NL	3L value set	**Mild depression**	0.63 (0.12)	-0.03	1	0.01 (0.003; 0.02)
	3L to 5L crosswalk		0.62 (0.14)	0.07	0.95	
US	3L value set		0.71 (0.15)	0.27	1	**0.08 (0.07; 0.08)**
	3L to 5L crosswalk		0.63 (0.15)	0.11	0.96	
JP	3L value set		0.65 (0.08)	0.42	1	-0.02 (-0.03; -0.02)
	3L to 5L crosswalk		0.67 (0.09)	0.42	0.92	
NL	3L value set	**Moderate depression**	0.57 (0.22)	-0.07	1	0.01 (0.001; 0.02)
	3L to 5L crosswalk		0.56 (0.15)	-0.03	0.95	
US	3L value set		0.67 (0.16)	0.21	1	**0.09 (0.09; 0.10)**
	3L to 5L crosswalk		0.58 (0.16)	-0.02	0.96	
JP	3L value set		0.62 (0.10)	0.15	1	-0.02 (-0.02; -0.01)
	3L to 5L crosswalk		0.64 (0.09)	0.35	0.92	

Country	Scoring method	Patient population	Mean utilities (SD)	Min	Max	3Lvs – 3L to 5L cw (95% CI)
NL	3L value set	**Severe depression**	0.30 (0.24)	-0.23	0.80	**-0.07 (-0.08; -0.06)**
	3L to 5L crosswalk		0.37(0.19)	-0.15	0.80	
US	3L value set		0.47 (0.19)	-0.01	0.84	**0.09 (0.08; 0.09)**
	3L to 5L crosswalk		0.38 (0.20)	-0.18	0.86	
JP	3L value set		0.50 (0.16)	-0.01	0.78	-0.03 (-0.04; -0.02)
	3L to 5L crosswalk		0.53 (0.11)	0.24	0.81	
NL	3L value set	**Mild low back pain**	0.79 (0.08)	0.43	1	0.06 (0.05; 0.06)
	3L to 5L crosswalk		0.73 (0.07)	0.50	0.95	
US	3L value set		0.79 (0.06)	0.51	1	0.06 (0.05; 0.07)
	3L to 5L crosswalk		0.73 (0.10)	0.38	0.96	
JP	3L value set		0.70 (0.06)	0.51	1	-0.04 (-0.04; -0.04)
	3L to 5L crosswalk		0.74 (0.07)	0.53	0.92	
NL	3L value set	**Moderate low back pain**	0.68 (0.19)	0.09	1	0.03 (0.02; 0.04)
	3L to 5L crosswalk		0.65 (0.13)	0.24	0.95	
US	3L value set		0.72 (0.14)	0.31	1	**0.09 (0.08; 0.09)**
	3L to 5L crosswalk		0.63 (0.14)	0.19	0.96	
JP	3L value set		0.65 (0.08)	0.42	1	-0.03 (-0.03; -0.03)
	3L to 5L crosswalk		0.68 (0.09)	0.44	0.92	
NL	3L value set	**Severe low back pain**	0.43 (0.25)	-0.11	0.81	-0.04 (-0.05; -0.02)
	3L to 5L crosswalk		0.47 (0.15)	0.01	0.74	
US	3L value set		0.54 (0.18)	0.08	0.82	**0.11 (0.09; 0.11)**
	3L to 5L crosswalk		0.43 (0.16)	0.001	0.77	
JP	3L value set		0.54 (0.09)	0.05	0.72	-0.04 (-0.04; -0.03)
	3L to 5L crosswalk		0.58 (0.08)	0.33	0.77	
NL	3L value set	**Mild osteoarthritis**	0.80 (0.09)	0.37	1	0.05 (0.05; 0.06)
	3L to 5L crosswalk		0.75 (0.08)	0.55	0.95	
US	3L value set		0.80 (0.09)	0.36	1	0.06 (0.05; 0.07)
	3L to 5L crosswalk		0.74 (0.10)	0.46	0.96	
JP	3L value set		0.71 (0.10)	0.30	1	-0.04 (-0.04; -0.03)
	3L to 5L crosswalk		0.76 (0.08)	0.55	0.92	
NL	3L value set	**Moderate osteoarthritis**	0.76 (0.09)	0.33	1	**0.08 (0.07; 0.08)**
	3L to 5L crosswalk		0.68 (0.07)	0.50	0.95	
US	3L value set		0.77 (0.07)	0.45	1	**0.11 (0.11; 0.12)**
	3L to 5L crosswalk		0.65 (0.10)	0.38	0.96	
JP	3L value set		0.66 (0.05)	0.51	1	-0.02 (-0.02; -0.02)
	3L to 5L crosswalk		0.68 (0.07)	0.52	0.92	
NL	3L value set	**Severe osteoarthritis**	0.52 (0.26)	-0.03	0.89	0.001 (-0.01; 0.01)
	3L to 5L crosswalk		0.52 (0.16)	0.07	0.85	
US	3L value set		0.61 (0.18)	0.27	0.85	**0.12 (0.11; 0.12)**
	3L to 5L crosswalk		0.49 (0.15)	0.09	0.83	
JP	3L value set		0.58 (0.08)	0.38	0.77	-0.02 (-0.02; -0.02)
	3L to 5L crosswalk		0.60 (0.78)	0.37	0.79	

Country	Scoring method	Patient population	Mean utilities (SD)	Min	Max	3Lvs – 3L to 5L cw (95% CI)
NL	3L value set	**Mild cancer**	0.92 (0.09)	0.69	1	0.05 (0.04; 0.05)
	3L to 5L crosswalk		0.87 (0.08)	0.62	0.95	
US	3L value set		0.91 (0.08)	0.77	1	0.02 (0.02; 0.03)
	3L to 5L crosswalk		0.89 (0.08)	0.62	0.96	
JP	3L value set		0.88 (0.12)	0.65	1	0.02 (0.01; 0.02)
	3L to 5L crosswalk		0.86 (0.07)	0.66	0.92	
NL	3L value set	**Moderate cancer**	0.73 (0.15)	0.21	1	0.03 (0.03; 0.04)
	3L to 5L crosswalk		0.70 (0.11)	0.38	0.95	
US	3L value set		0.76 (0.11)	0.42	1	0.06 (0.05; 0.07)
	3L to 5L crosswalk		0.70 (0.13)	0.39	0.96	
JP	3L value set		0.69 (0.09)	0.45	1	-0.03 (-0.03; -0.02)
	3L to 5L crosswalk		0.72 (0.09)	0.50	0.92	
NL	3L value set	**Severe cancer**	0.55 (0.40)	-0.33	1	0.04 (0.03; 0.04)
	3L to 5L crosswalk		0.51 (0.36)	-0.31	1	
US	3L value set		0.62 (0.34)	-0.11	1	**0.13 (0.12; 0.14)**
	3L to 5L crosswalk		0.49 (0.40)	-0.42	0.96	
JP	3L value set		0.56 (0.34)	-0.11	1	-0.04 (-0.05; -0.02)
	3L to 5L crosswalk		0.60 (0.24)	0.10	0.92	

3Lvs: EQ-5D-3L value set. cw: crosswalk. NL: the Netherlands. US: United States. JP: Japan. CI: confidence interval. Differences in utilities between 3L value set and 3L to 5L crosswalk ≥0.074 (i.e., the minimally clinically important difference) are highlighted in bold.

For the Netherlands, differences were clinically relevant in 2 out of 12 patient populations (i.e., 17%), for the United States in 8 out of 12 (i.e., 67%), for Japan no clinically relevant differences were found. Note that only 12 possible comparisons could be done as no treatment effect was simulated at baseline. That is, four health conditions times three severity levels, also referred to as patient population.

Table 3 | Differences in utilities estimated by 5L value sets and 5L to 3L crosswalks

Country	Scoring method	Patient population	Mean utilities (SD)	Min	Max	3Lvs – 3L to 5L cw (95% CI)
NL	5L value set	**Mild depression**	0.66 (0.26)	-0.29	1	-0.03 (-0.03; -0.01)
	5L to 3L crosswalk		0.69 (0.20)	0.003	1	
US	5L value set		0.70 (0.28)	-0.37	1	-0.06 (-0.08; -0.05)
	5L to 3L crosswalk		0.76 (0.15)	0.20	1	
JP	5L value set		0.71 (0.16)	0.13	1	0.01 (0.005; 0.01)
	5L to 3L crosswalk		0.70 (0.13)	0.30	1	
NL	5L value set	**Moderate depression**	0.58 (0.29)	-0.41	1	-0.04 (-0.05; -0.03)
	5L to 3L crosswalk		0.62 (0.23)	-0.16	1	
US	5L value set		0.62 (0.31)	-0.45	1	**-0.10 (-0.12; -0.08)**
	5L to 3L crosswalk		0.72 (0.17)	0.13	1	
JP	5L value set		0.67 (0.17)	0.08	1	-0.002 (-0.01; 0.003)
	5L to 3L crosswalk		0.67 (0.13)	0.24	1	

Country	Scoring method	Patient population	Mean utilities (SD)	Min	Max	3Lvs – 3L to 5L cw (95% CI)
NL	5L value set	**Severe depression**	0.37 (0.37)	-0.41	1	**-0.08 (-0.10; -0.08)**
	5L to 3L crosswalk		0.45 (0.28)	-0.26	1	
US	5L value set		0.40 (0.40)	-0.45	1	-0.20 (-0.22; -0.18)
	5L to 3L crosswalk		0.60 (0.22)	-0.04	1	
JP	5L value set		0.55 (0.21)	0.07	1	-0.04 (-0.04; -0.03)
	5L to 3L crosswalk		0.59 (0.17)	-0.06	1	
NL	5L value set	**Mild low back pain**	0.45 (0.34)	-0.18	0.80	**-0.10 (-0.11; -0.09)**
	5L to 3L crosswalk		0.55 (0.22)	0.17	0.81	
US	5L value set		0.39 (0.35)	-0.22	0.78	**-0.22 (-0.24; -0.20)**
	5L to 3L crosswalk		0.61 (0.16)	0.35	0.81	
JP	5L value set		0.52 (0.18)	0.24	0.76	-0.03 (-0.04; -0.04)
	5L to 3L crosswalk		0.55 (0.10)	0.41	0.69	
NL	5L value set	**Moderate low back pain**	0.42 (0.31)	-0.28	0.86	**-0.10 (-0.12; -0.09)**
	5L to 3L crosswalk		0.52 (0.20)	-0.11	0.84	
US	5L value set		0.37 (0.32)	-0.32	0.90	**-0.22 (-0.24; -0.20)**
	5L to 3L crosswalk		0.59 (0.15)	0.06	0.83	
JP	5L value set		0.52 (0.18)	0.13	0.87	-0.03 (-0.04; -0.02)
	5L to 3L crosswalk		0.54 (0.10)	0.005	0.77	
NL	5L value set	**Severe low back pain**	0.24 (0.22)	-0.08	0.75	**-0.18 (-0.18; -0.16)**
	5L to 3L crosswalk		0.42 (0.13)	0.27	0.72	
US	5L value set		0.18 (0.23)	-0.15	0.65	**-0.33 (-0.34; -0.31)**
	5L to 3L crosswalk		0.51 (0.09)	0.39	0.72	
JP	5L value set		0.45 (0.15)	0.25	0.71	-0.05 (-0.06; -0.04)
	5L to 3L crosswalk		0.50 (0.07)	0.42	0.63	
NL	5L value set	**Mild osteoarthritis**	0.82 (0.17)	0.05	1	-0.001 (-0.01; 0.004)
	5L to 3L crosswalk		0.82 (0.13)	0.32	1	
US	5L value set		0.82 (0.20)	-0.02	1	-0.006 (-0.02; 0.01)
	5L to 3L crosswalk		0.83 (0.11)	0.44	1	
JP	5L value set		0.82 (0.15)	0.33	1	0.06 (0.05; 0.06)
	5L to 3L crosswalk		0.76 (0.13)	0.44	1	
NL	5L value set	**Moderate osteoarthritis**	0.78 (0.13)	-0.08	1	-0.002 (-0.01; 0.002)
	5L to 3L crosswalk		0.78 (0.10)	0.20	1	
US	5L value set		0.75 (0.15)	-0.06	1	-0.03 (-0.04; -0.03)
	5L to 3L crosswalk		0.79 (0.09)	0.38	1	
JP	5L value set		0.75 (0.12)	0.30	1	0.06 (0.05; 0.06)
	5L to 3L crosswalk		0.69 (0.10)	0.43	1	
NL	5L value set	**Severe osteoarthritis**	0.59 (0.35)	-0.38	0.89	-0.05 (-0.07; -0.04)
	5L to 3L crosswalk		0.65 (0.27)	-0.23	0.87	
US	5L value set		0.55 (0.38)	-0.55	0.94	**-0.13 (-0.15; -0.11)**
	5L to 3L crosswalk		0.68 (0.23)	-0.07	0.86	
JP	5L value set		0.63 (0.23)	-0.001	0.90	0.03 (0.03; 0.04)
	5L to 3L crosswalk		0.60 (0.19)	-0.09	0.81	

Country	Scoring method	Patient population	Mean utilities (SD)	Min	Max	3Lvs – 3L to 5L cw (95% CI)
NL	5L value set	**Mild cancer**	0.85 (0.20)	-0.10	1	-0.01 (-0.01; -0.003)
	5L to 3L crosswalk		0.86 (0.18)	-0.02	1	
US	5L value set		0.85 (0.22)	-0.18	1	-0.03 (-0.04; -0.02)
	5L to 3L crosswalk		0.88 (0.15)	0.23	1	
JP	5L value set		0.85 (0.18)	0.20	1	0.01 (0.01; 0.02)
	5L to 3L crosswalk		0.83 (0.18)	0.31	1	
NL	5L value set	**Moderate cancer**	0.76 (0.26)	-0.34	1	-0.03 (-0.03; -0.02)
	5L to 3L crosswalk		0.79 (0.20)	-0.06	1	
US	5L value set		0.75 (0.28)	-0.42	1	-0.06 (-0.07; -0.05)
	5L to 3L crosswalk		0.81 (0.16)	0.18	1	
JP	5L value set		0.76 (0.21)	0.10	1	0.01 (0.01; 0.02)
	5L to 3L crosswalk		0.75 (0.19)	0.27	1	
NL	5L value set	**Severe cancer**	0.55 (0.50)	-0.45	1	-0.06 (-0.08; -0.05)
	5L to 3L crosswalk		0.61 (0.41)	-0.33	1	
US	5L value set		0.52 (0.53)	-.57	1	**-0.16 (-0.19; -0.14)**
	5L to 3L crosswalk		0.69 (0.32)	-.11	1	
JP	5L value set		0.65 (0.33)	-.02	1	-0.01 (-0.02; -0.005)
	5L to 3L crosswalk		0.66 (0.30)	-.11	1	

5Lvs: EQ-5D-5L value set. cw: crosswalk. NL: the Netherlands. US: United States. JP: Japan. NL: the Netherlands. US: United States. JP: Japan. CI: confidence interval. Differences in utilities between 5L value set and 5L to 3L crosswalk ≥ 0.074 (i.e., the minimally clinically important difference) are highlighted in bold. For the Netherlands, differences were clinically relevant in 4 out of 12 patient populations (i.e., 33%), for the United States in 6 out of 12 (i.e., 50%), for Japan no clinically relevant differences were found. Note that only 12 possible comparisons could be done as no treatment effect was simulated at baseline. That is, four health conditions times three severity levels, also referred to as patient population.

QALYs

Differences in QALYs between EQ-5D value sets and crosswalks ranged from -0.18 (i.e., between the US 5L value set and 5L to 3L crosswalk, Table 4, scenario 16) to 0.13 (i.e., between the US 3L value set and 3L to 5L crosswalk, Table 4, scenario 26). QALYs statistically significantly differed between EQ-5D value sets and crosswalks in all 36 scenarios for the three countries. No clinically relevant differences between the 3L value set and 3L to 5L crosswalk were found for Japan and the Netherlands, whereas differences were clinically relevant in 14% of scenarios for the US. Clinically relevant differences between the 5L value set and 5L to 3L crosswalk were found in 8%, 25%, and 50% of scenarios, for the Netherlands, Japan, and the United States, respectively.

Table 4 | Overview of differences in QALY between EQ-5D value sets and crosswalks

Country	Scoring method	Scenario	Patient population	Effect size	QALYs (SD)	Min	Max	QALY – QALY cw (95% CI)
NL	3L value set	(7)	**Severe depression**	Small	0.44 (0.20)	-0.07	0.88	**-0.07 (-0.07; -0.06)**
	3L to 5L crosswalk				0.51 (0.15)	0.06	0.84	
US	3L value set				0.56 (0.15)	0.20	0.91	0.04 (0.04; 0.05)
	3L to 5L crosswalk				0.52 (0.15)	0.10	0.85	
JP	3L value set				0.58 (0.12)	0.25	0.91	-0.05 (-0.05; -0.04)
	3L to 5L crosswalk				0.63 (0.08)	0.41	0.81	
NL	3L value set	(26)	**Severe osteoarthritis**	Medium	0.28 (0.13)	-0.03	0.46	-0.05 (-0.06; -0.05)
	3L to 5L crosswalk				0.33 (0.08)	0.03	0.50	
US	3L value set				0.45 (0.09)	0.24	0.60	**0.13 (0.12; 0.13)**
	3L to 5L crosswalk				0.33 (0.08)	-0.02	0.53	
JP	3L value set				0.52 (0.09)	0.30	0.65	0.002 (-0.01; 0.002)
	3L to 5L crosswalk				0.52 (0.04)	0.33	0.63	
NL	5L value set	(16)	**Severe low back pain**	Small	0.42 (0.18)	-0.06	0.87	-0.07 (-0.08; -0.07)
	5L to 3L crosswalk				0.49 (0.15)	0.12	0.86	
US	5L value set				0.39 (0.18)	-0.15	0.82	**-0.18 (-0.19; -0.18)**
	5L to 3L crosswalk				0.57 (0.12)	0.23	0.86	
JP	5L value set				0.43 (0.15)	0.05	0.82	-0.15 (-0.15; -0.14)
	5L to 3L crosswalk				0.58 (0.10)	0.21	0.81	
NL	5L value set	(22)	**Moderate osteoarthritis**	Small	0.71 (0.15)	0.13	0.96	0.02 (0.02; 0.02)
	5L to 3L crosswalk				0.69 (0.14)	0.25	0.94	
US	5L value set				0.69 (0.15)	0.13	0.97	-0.03 (-0.03; -0.02)
	5L to 3L crosswalk				0.72 (0.11)	0.37	0.93	
JP	5L value set				0.73 (0.12)	0.22	0.97	**0.05 (0.04; 0.05)**
	5L to 3L crosswalk				0.68 (0.10)	0.34	0.91	

3Lvs: EQ-5D-3L value set. 5Lvs: EQ-5D-5L value set. cw: crosswalk. NL: the Netherlands. US: United States. JP: Japan. CI: confidence interval.

Scenario 7 represents the lowest difference in QALYs between 3L value sets and 3L to 5L crosswalks across all scenarios (i.e., -0.07 in bold). Scenario 26 represents the largest difference in incremental QALYs between 3L value sets and 3L to 5L crosswalks across all scenarios (i.e., 0.13 in bold).

Scenario 16 represents the lowest difference in QALYs between 5L value sets and 5L to 3L crosswalks across all scenarios (i.e., -0.18 in bold). Scenario 22 represents the largest difference in incremental QALYs between 5L value sets and 5L to 3Lcrosswalks across all scenarios (i.e., 0.05 in bold).

Cost-utility analysis

Incremental QALYs

Over all scenarios, the largest difference in incremental QALYs between 3L value sets and 3L to 5L crosswalks was 0.06 using Dutch valuations (Table 5, scenario 9), while the largest difference between 5L value sets and 5L to 3L crosswalks was 0.08 using US valuations (Table 6, scenario 33).

ICER

The largest differences in ICERs between crosswalks and EQ-5D value sets were found in scenarios with small effect sizes, particularly those with mild health states regardless to the health condition (Table 5, scenarios 1 and 19; Table 6 scenarios 1, 19, 28, 31). Depending on the country, the magnitude of the difference in ICERs was so large that it could in turn impact the decision of whether an intervention is cost-effective or not (i.e., whether the ICER lies below a country's WTP per QALY gained). For example, in the scenario 1, ICERs estimated by 3L to 5L crosswalk, and the Japanese 3L value set differed tremendously, with the biggest difference being €11,063/QALY gained for the 3L to 5L crosswalk and €855,681/QALY gained for the Japanese 3L value set (Appendix 3). The differences in ICERs were generally larger for the EQ-5D-3L compared with the EQ-5D-5L and were most pronounced for Japan. Detailed information on ICERs can be found in Appendix 4.

Probabilities of cost-effectiveness

Larger differences between crosswalks and EQ-5D value sets were found in scenarios with small treatment effect sizes, while this was less evident for scenarios with medium and large ones. For example, for small effect sizes, at a WTP of €20,000/QALY gained, the largest differences in the probability of cost-effectiveness between EQ-5D value sets and crosswalks were found for mild depression (difference between 3L value set and 3L to 5L crosswalk=0.42, Table 5, scenario 1) and moderate cancer (difference between 5L value set and 5L to 3L crosswalk=0.63, Table 6, scenario 31) using Japanese valuations. For medium effect sizes, at the same WTP threshold, the largest differences were found for mild cancer (difference between 3L value set and 3L to 5L crosswalk= 0.06, Table 5, scenario 29) and for severe low back pain (difference between 5L value set and 5L to 3L crosswalk=0.01, Table 6, scenario 17) using Japanese valuations. For large effect sizes, the largest difference was found for mild osteoarthritis using Japanese valuations (difference between 3L value set and 3L to 5L crosswalk=0.08, Table 5, scenario 21) and no differences were found between 5L value sets and 5L to 3L crosswalks. At a WTP of €50,000/QALY gained, the largest differences were found in scenarios including small effect sizes for mild depression (difference between 3L value set and 3L to 5L crosswalk=0.47, Table 5, scenario 1) and moderate cancer (difference between 5L value set and 5L to 3L crosswalk=0.54, Table 6, scenario 31) using Japanese valuations, while no differences were found in all scenarios with medium and large effect sizes, except for severe osteoarthritis using Dutch valuations (difference between 3L value set and 3L to 5L crosswalk=0.01, Table 5, scenario 26).

Table 5 | Overview of cost-utility outcomes: Differences between 3L value sets and 3L to 5L crosswalks

Country	Scoring method	Scenario	Patient population	Effect size	Δ utilities 3Lvs – 3L to 5L cw (95% CI)	Δ QALYs 3Lvs – 3L to 5L cw (95% CI)	Δ incremental QALYs 3Lvs – 3L to 5L cw	Δ ICER, €/point 3Lvs – 3L to 5L cw	Δ Probability of cost-effectiveness			
									p_{CE} (0)	p_{CE} (20,000)	p_{CE} (30,000)	p_{CE} (50,000)
NL	3L value set	(1)	**Mild depression**	Small	0.01 (0.003; 0.02)	-0.03 (-0.03; -0.02)	0.02	-2628	0	0.10	0.07	0.03
	3L to 5L crosswalk											
US	3L value set				0.08 (0.07; 0.08)	0.04 (0.03; 0.04)	0.01	-5190	0	**0.15**	**0.15**	**0.13**
	3L to 5L crosswalk											
JP	3L value set				-0.02 (-0.03; -0.02)	-0.01 (-0.01; 0.001)	0.02	-844618	0	**0.42**	**0.47**	**0.47**
	3L to 5L crosswalk											
NL	3L value set	(9)	**Severe depression**	Large	-0.07 (-0.08; -0.06)	-0.04 (-0.05; -0.04)	**0.06**	-169	0	0	0	0
	3L to 5L crosswalk											
US	3L value set				0.09 (0.08; 0.09)	0.06 (0.06; 0.07)	0.004	-16	0	0	0	0
	3L to 5L crosswalk											
JP	3L value set				-0.03 (-0.04; -0.02)	-0.03 (-0.04; -0.02)	0.03	-266	0	0	0	0
	3L to 5L crosswalk											
NL	3L value set	(11)	**Mild low back pain**	Medium	0.06 (0.05; 0.06)	-0.01 (-0.01; 0.001)	0.02	-761	0	0	0	0
	3L to 5L crosswalk											
US	3L value set				0.06 (0.05; 0.07)	0.03 (0.03; 0.04)	-0.01	532	0	0	0	0
	3L to 5L crosswalk											
JP	3L value set				-0.04 (-0.04; -0.04)	0.004 (0.0004; 0.01)	0.003	-251	0	0	0	0
	3L to 5L crosswalk											
NL	3L value set	(19)	**Mild osteo-arthritis**	Small	0.05 (0.05; 0.06)	0.05 (0.05; 0.05)	0.004	-533	0	-0.01	-0.02	-0.03
	3L to 5L crosswalk											
US	3L value set				0.06 (0.05; 0.07)	0.05 (0.05; 0.06)	0.01	-3176	0	**0.14**	**0.13**	**0.11**
	3L to 5L crosswalk											
JP	3L value set				-0.04 (-0.04; -0.03)	0.02 (0.01; 0.02)	0.02	-38836	0	**0.37**	**0.15**	**0.15**
	3L to 5L crosswalk											

Country	Scoring method	Scenario	Patient population	Effect size	Δ utilities 3Lvs – 3L to 5L cw (95% CI)	Δ QALYs 3Lvs – 3L to 5L cw (95% CI)	Δ incremental QALYs 3Lvs – 3L to 5L cw	Δ ICER, €/point 3Lvs – 3L to 5L cw	Δ Probability of cost-effectiveness			
									p_{CE} (0)	**p_{CE} (20,000)**	**p_{CE} (30,000)**	**p_{CE} (50,000)**
NL	3L value set	(21)	**Mild osteo-arthritis**	Large	0.05 (0.05; 0.06)	0.05 (0.05; 0.06)	0.002	-211	0	-0.001	0	0
	3L to 5L crosswalk											
US	3L value set				0.06 (0.05; 0.07)	0.05 (0.05; 0.06)	0.01	-2161	0	0.05	0.002	0
	3L to 5L crosswalk											
JP	3L value set				-0.04 (-0.04; -0.03)	0.04 (0.04; 0.05)	0.05	-5954	0	**0.08**	0.03	0.001
	3L to 5L crosswalk											
NL	3L value set	(23)	**Moderate osteo-arthritis**	Medium	0.08 (0.07; 0.08)	0.001 (-0.03; 0.01)	0.01	-499	0	-0.002	-0.001	0
	3L to 5L crosswalk											
US	3L value set				0.11 (0.11; 0.12)	0.05 (0.05; 0.06)	-0.02	1708	0	-0.01	-0.001	0
	3L to 5L crosswalk											
JP	3L value set				-0.02 (-0.02; -0.02)	0.02 (0.02; 0.03)	0.002	-240	0	-0.002	-0.001	0
	3L to 5L crosswalk											
NL	3L value set	(29)	**Mild cancer**	Medium	0.05 (0.04; 0.05)	0.01 (0.01; 0.02)	0.003	-132	0	0.01	0.003	0
	3L to 5L crosswalk											
US	3L value set				0.02 (0.02; 0.03)	0.01 (0.01; 0.01)	0.01	-455	0	0.03	0.01	0.001
	3L to 5L crosswalk											
JP	3L value set				0.02 (0.01; 0.02)	0.02 (0.02; 0.02)	0.03	-1085	0	**0.06**	0.02	0.003
	3L to 5L crosswalk											

Δ: differences. 3Lvs: EQ-5D-3L value set. cw: crosswalk. NL: the Netherlands. US: United States. JP: Japan. CI: confidence interval. ICER: Incremental cost-effectiveness ratio. pCE(0): probability of cost-effectiveness at a zero willingness-to-pay per QALY gained. pCE(10,000): probability of cost-effectiveness at a willingness-to-pay per QALY gained of 10,000 euros.

Scenario 9 represents the largest difference in incremental QALYs between 3L value sets and 3L to 5L crosswalks across all scenarios (i.e., 0.06 in bold).

Scenarios 1, 11, 19, 21, and 29 were presented to illustrate the impact of crosswalks on ICERs and probabilities of cost-effectiveness for small, medium, and large treatment effect sizes.

Table 6 | Overview of cost-utility outcomes: Differences between 5L value sets and 5L to 3L crosswalks

Country	Scoring method	Scenario	Patient population	Effect size	Δ utilities 5Lvs – 5L to 3L cw (95% CI)	Δ QALYs 5Lvs – 5L to 3L cw (95% CI)	Δ incremental QALYs 5Lvs – 5L to 3L cw	Δ ICER, €/point 5Lvs – 5L to 3L cw	Δ Probability of cost-effectiveness			
									p_{CE} (0)	p_{CE} (20,000)	p_{CE} (30,000)	p_{CE} (50,000)
NL	5L value set	(1)	**Mild depression**	Small	-0.03 (-0.03; -0.01)	0.001 (-0.003; 0.005)	0.001	-385	0	-0.003	0.001	-0.01
	5L to 3L crosswalk											
US	5L value set				-0.06 (-0.08; -0.05)	-0.05 (-0.05; -0.04)	0.01	-2226	0	0.05	0.02	0.001
	5L to 3L crosswalk											
JP	5L value set				0.01 (0.005; 0.01)	0.01 (0.01; 0.02)	-0.01	7598	0	-0.25	-0.26	-0.16
	5L to 3L crosswalk											
NL	5L value set	(17)	**Severe low back pain**	Medium	-0.18 (-0.18; -0.16)	-0.07 (-0.08; -0.07)	0.02	-724	0	0	0	0
	5L to 3L crosswalk											
US	5L value set				-0.33 (-0.34; -0.31)	-0.18 (-0.19; -0.18)	0.03	-1741	0	0.003	0	0
	5L to 3L crosswalk											
JP	5L value set				-0.05 (-0.06; -0.04)	-0.15 (-0.15; -0.14)	0.01	-1315	0	0.01	0	0
	5L to 3L crosswalk											
NL	5L value set	(19)	**Mild osteo-arthritis**	Small	-0.001 (-0.01; 0.004)	0.01 (0.01; 0.02)	-0.002	181	0	-0.03	-0.03	-0.03
	5L to 3L crosswalk											
US	5L value set				-0.006 (-0.02; 0.01)	-0.01 (-0.02; -0.01)	0.01	-980	0	0.04	0.02	0.01
	5L to 3L crosswalk											
JP	5L value set				0.06 (0.05; 0.06)	0.04 (0.04; 0.05)	0.004	-644	0	0.001	-0.02	-0.02
	5L to 3L crosswalk											
NL	5L value set	(24)	**Moderate osteo-arthritis**	Large	-0.002 (-0.01; 0.002)	0.01 (0.01; 0.01)	-0.01	-206	0	0	0	0
	5L to 3L crosswalk											
US	5L value set				-0.03 (-0.04; -0.03)	-0.02 (-0.03; -0.02)	0.02	390	0	0	0	0
	5L to 3L crosswalk											
JP	5L value set				0.06 (0.05; 0.06)	0.04 (0.04; 0.05)	0.01	165	0	0	0	0
	5L to 3L crosswalk											

Country	Scoring method	Scenario	Patient population	Effect size	Δ utilities	Δ QALYs	Δ incremental QALYs	Δ ICER, €/point	Δ Probability of cost-effectiveness			
					5Lvs – 5L to 3L cw (95% CI)	5Lvs – 5L to 3L cw (95% CI)	5Lvs – 5L to 3L cw	5Lvs – 5L to 3L cw	p_{CE} (0)	p_{CE} (20,000)	p_{CE} (30,000)	p_{CE} (50,000)
NL	5L value set	(28)	**Mild cancer**	Small	-0.01	0.01	0.003	-4262	0	0.05	0.05	0.07
	5L to 3L crosswalk				(-0.01; -0.003)	(0.001; 0.01)						
US	5L value set				-0.03	-0.03	0.01	-14380	0	0.18	0.18	0.15
	5L to 3L crosswalk				(-0.04; -0.02)	(-0.04; -0.02)						
JP	5L value set				0.01	0.02	-0.01	3523	0	-0.12	-0.12	-0.10
	5L to 3L crosswalk				(0.01; 0.02)	(0.02; 0.03)						
NL	5L value set	(31)	**Moderate cancer**	Small	-0.03	0.001	-0.02	1746	0	-0.11	-0.08	-0.06
	5L to 3L crosswalk				(-0.03; -0.02)	(-0.003; 0.01)						
US	5L value set				-0.06	-0.04	0.01	-1799	0	0.03	-0.0003	-0.01
	5L to 3L crosswalk				(-0.07; -0.05)	(-0.05; -0.04)						
JP	5L value set				0.01	0.01	0.06	-66527	0	0.63	0.59	0.54
	5L to 3L crosswalk				(0.01; 0.02)	(0.01; 0.02)						
NL	5L value set	(33)	**Moderate cancer**	Large	-0.03	-0.03	0.03	96	0	0	0	0
	5L to 3L crosswalk				(-0.03; -0.02)	(-0.03; -0.02)						
US	5L value set				-0.06	-0.07	0.08	354	0	0	0	0
	5L to 3L crosswalk				(-0.07; -0.05)	(-0.08; -0.07)						
JP	5L value set				0.01	0.001	-0.001	-17	0	0	0	0
	5L to 3L crosswalk				(0.01; 0.02)	(-0.005; 0.01)						

Δ: differences. 5Lvs: EQ-5D-5L value set. cw: crosswalk. NL: the Netherlands. US: United States. JP: Japan. CI: confidence interval. ICER: Incremental cost-effectiveness ratio. pCE(0): probability of cost-effectiveness at a zero willingness-to-pay per QALY gained. pCE(10,000): probability of cost-effectiveness at a willingness-to-pay per QALY gained of 10,000 euros.

Scenario 33 represents the largest difference in incremental QALYs between 5L value sets and 5L to 3L crosswalks across all scenarios (i.e., 0.08 in bold).

Scenarios 1, 17, 19, 24 28 and 31 were presented to illustrate the impact of crosswalks on ICERs and probabilities of cost-effectiveness for small, medium, and large treatment effect sizes.

Discussion

Main findings

The aim of the current study was to evaluate the impact of using crosswalks or EQ-5D value sets on reimbursement decisions in a wide variety of simulated trial-based economic evaluations for the Netherlands, the United States, and Japan. Results showed that differences exist in means and distributions of utilities, incremental QALYs, and ICER point estimates between scoring methods in all simulated scenarios and countries. In our study, this only affected reimbursement decisions in scenarios with small treatment effect sizes, especially in mild health states regardless of the health condition. This impact was more pronounced in the United States and Japan than in the Netherlands. In scenarios with medium and large effect sizes, the impact on the probability of cost-effectiveness was relatively small in all countries. Our findings suggest that caution is warranted when using crosswalks, especially when treatment effect sizes are small and in countries that were not included in the crosswalk development studies (i.e., all countries except Denmark, England, Italy, the Netherlands, Poland, and Scotland).

Interpretation of the findings and comparison with the literature

In line with previous studies,[18,24–29] our study found that different EQ-5D scoring methods resulted in different utilities estimates, which in turn resulted in different incremental QALY and ICER estimates. Differences in utilities and QALYs between EQ-5D scoring methods in certain scenarios and conditions may be due to differences in utility decrements between health dimensions in the different value sets but also to the probability of observing certain response levels within conditions (e.g., low back pain patients have a high probability of scoring severe response levels on the "pain/ discomfort" dimension). The magnitude of the differences and their clinical relevance differed across countries, with differences generally being larger in the United States and Japan than in the Netherlands.

A previous study concluded that there was no impact on reimbursement decisions of the scoring method used.[29] In contrast, we now show that in some scenarios, particularly those with small treatment effect sizes, the use of crosswalks instead of country-specific EQ-5D value sets impacts cost-utility outcomes to such an extent that this may influence reimbursement decisions. The difference in findings and conclusion between our previous and current study may be explained by the fact that the interventions of the case studies used in our previous study were on average "less effective" and "more costly" than control. In the present study, we simulated scenarios with interventions that were "more effective" and "more costly", which is a more likely scenario to occur in real-life reimbursement decisions. Our current findings also show that different EQ-5D scoring approaches were more likely to impact a reimbursement decision for countries that were not used in the development of the crosswalk. This may be due to the fact that the sample included in the crosswalk development study may not represent the preferences of other populations, particularly those with considerably different views on health-related quality of life.

Strengths and limitations

One of the strengths of this study is that the impact on cost-utility outcomes was evaluated for three different countries, two of which were not used for the development of the crosswalk and differed considerably from the Dutch value set in terms of the utility decrements assigned to the different health dimensions of the EQ-5D.[20,22] Another strength is our use of simulated data and a wide range of scenarios. These scenarios were based on empirical studies in chronic health conditions that have a high impact on populations' health-related quality of life and/or life expectancy. Moreover, the simulated scenarios included different severity levels of the included health conditions and interventions with small, medium, and large impacts on health-related quality of life. Furthermore, full trial-based economic evaluations were performed including the assessment of uncertainty around ICER estimates.

A limitation of this study is that cost data were simulated in such a way that cost differences were not statistically significant, but we do not expect this to change our overall conclusion that caution is warranted when using crosswalks for estimating EQ-5D utilities, particularly when effect sizes are small. Additionally, only three countries were investigated, whereas EQ-5D value sets are available for many countries. However, we deliberately chose countries with considerably different utility decrements to include the full spectrum of preferences from other countries.

Recommendations for research and practice

The current results indicate that the use of crosswalks may impact on reimbursement decisions in situations where treatment effect sizes are small, and interventions are more costly compared to control. Given the rigorous quality control protocols for the EQ-5D valuation studies, the most appropriate EQ-5D scoring method is the available country-specific value set developed using the most recent version of the EQ-VT protocol.[15] In case of multi-country randomized clinical-trials, researchers are recommended to check the HTA guidelines of the participating countries for the most appropriate choice. Nonetheless, there are cases in which the decision on which value set to use is more complex, such as when a value set is only available for one of the EQ-5D version, while data have been collected using the other version of the EQ-5D. In such situations, caution is needed when using crosswalks as they may impact cost-utility outcomes, particularly in countries that were not included in the developments of the crosswalks. For further details and guidance about the choice of scoring methods, researchers are advised to check EuroQol recommendations.[23]

It is important to note that health economic models submitted to HTA agencies rarely use directly measured utilities, and that there is considerable freedom in which utilities are used. Thus, the finding of this study that there are considerable differences between the different valuation approaches do not necessarily result in an impact on QALY estimates in these models.

Conclusions

Crosswalks may be used when value sets are missing for a specific country or jurisdiction. However, our findings indicate that reimbursement decisions may change in situations with small effect sizes and countries that were not included in the development of the crosswalks. Therefore, when EQ-5D value sets are not available, researchers and decision-makers should be aware that the use of crosswalk is likely to impact decisions.

References

1. Brazier J, Ara R, Rowen D, Chevrou-Severac H. A Review of Generic Preference-Based Measures for Use in Cost-Effectiveness Models. *PharmacoEconomics*. 2017 Dec 1; **35**(1): 21–31.
2. Zhou T, Guan H, Wang L, Zhang Y, Rui M, Ma A. Health-Related Quality of Life in Patients With Different Diseases Measured With the EQ-5D-5L: A Systematic Review. *Frontiers in Public Health*. 2021; **9**: 802.
3. Finch AP, Brazier JE, Mukuria C. What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. *Eur J Health Econ*. 2017 May 30; **19**(4): 1–14.
4. Feng YS, Kohlmann T, Janssen MF, Buchholz I. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. *Qual Life Res*. 2021; **30**(3): 647–73.
5. Qian X, Tan RLY, Chuang LH, Luo N. Measurement Properties of Commonly Used Generic Preference-Based Measures in East and South-East Asia: A Systematic Review. *PharmacoEconomics*. 2020 Feb 1; **38**(2): 159–70.
6. EuroQol Group. EuroQol-a new facility for the measurement of health-related quality of life. *Health policy*. 1990; **16(**3): 199–208.
7. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011 Dec; **20**(10): 1727–36.
8. van Dongen JM, Jornada Ben Â, Finch AP, Rossenaar MMM, Biesheuvel-Leliefeld KEM, Apeldoorn AT, et al. Assessing the Impact of EQ-5D Country-specific Value Sets on Cost-utility Outcomes. *Medical Care*. 2021 Jan; **59**(1): 82–90.
9. Buchholz I, Janssen MF, Kohlmann T, Feng YS. A Systematic Review of Studies Comparing the Measurement Properties of the Three-Level and Five-Level Versions of the EQ-5D. *PharmacoEconomics*. 2018 Mar **23**; 1–17.
10. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Qual Life Res*. 2013; **22**(7): 1717–27.
11. Janssen MF, Bonsel GJ, Luo N. Is EQ-5D-5L Better Than EQ-5D-3L? A Head-to-Head Comparison of Descriptive Systems and Value Sets from Seven Countries. *Pharmacoeconomics*. 2018 Jun; **36**(6): 675–97.
12. Devlin N, Parkin D, Janssen B. Analysis of EQ-5D Values. In: Devlin N, Parkin D, Janssen B, editors. Methods for Analysing and Reporting EQ-5D Data [Internet]. Cham: Springer International Publishing; 2020 [cited 2021 Dec 9]. p. 61–86. Available from: https://doi.org/10.1007/978-3-030-47622-9_4
13. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. *Br Med Bull*. 2010 Dec 1; **96**(1): 5–21.
14. Devlin N, Finch AP, Parkin D. Guidance to Users of EQ-5D-5L Value Sets. In: Devlin N, Roudijk B, Ludwig K, editors. Value Sets for EQ-5D-5L: A Compendium, Comparative Review & User Guide [Internet]. Cham: Springer International Publishing; 2022 [cited 2022 Sep 15]. p. 213–33. Available from: https://doi.org/10.1007/978-3-030-89289-0_5
15. Stolk E, Ludwig K, Rand K, Hout B van, Ramos-Goñi JM. Overview, Update, and Lessons Learned From the International EQ-5D-5L Valuation Work: Version 2 of the EQ-5D-5L Valuation Protocol. *Value in Health*. 2019 Jan 1; 22(1): 23–30.
16. NICE. NICE health technology evaluations: the manual [Internet]. 2022 p. 181. Available from: https://www.nice.org.uk/process/pmg36
17. Alava MH, Pudney S, Wailoo A. The EQ-5D-5L Value Set for England: Findings of a Quality Assurance Program. *Value in Health*. 2020 May 1; **23**(5): 642–8.
18. Hernandez-Alava M, Pudney S. Econometric modelling of multiple self-reports of health states: The switch from EQ-5D-3L to EQ-5D-5L in evaluating drug therapies for rheumatoid arthritis. *Journal of Health Economics*. 2017 Sep 1; **55**: 139–52.

19. Hernández Alava M, Pudney S, Wailoo A. Estimating the relationship between EQ-5D-5L and EQ-5D-3L: results from an English Population Study [Internet]. University of Sheffield & University of York; 2020 [cited 2022 Jun 6]. Available from: https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/estimating-the-relationship-betweenE-Q-5D-5L-and-EQ-5D-3L.pdf

20. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, et al. Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets. Value in Health. 2012 Jul;15(5):708–15.

21. Hernandez-Alava M, Wailoo A, Pudney S. Methods for mapping between the EQ-5D-5L and the 3L for technology appraisal. 2017; **35**.

22. van Hout BA, Shaw JW. Mapping EQ-5D-3L to EQ-5D-5L. Value in Health [Internet]. 2021 May 17 [cited 2021 Jun 10];0(0). Available from: https://www.valueinhealthjournal.com/article/S1098-3015(21)00170-4/abstract

23. Devlin N, Finch AP, Parkin D. Guidance to users of EQ-5D-5L value sets. Forthcoming. In: Methods for Analysing and Reporting EQ-5D Data. 2021.

24. Mulhern B, Feng Y, Shah K, Janssen MF, Herdman M, Hout B van, et al. Comparing the UK EQ-5D-3L and English EQ-5D-5L Value Sets. *PharmacoEconomics*. 2018 Feb 23; **36**: 699–713.

25. Camacho EM, Shields G, Lovell K, Coventry PA, Morrison AP, Davies LM. A (five-) level playing field for mental health conditions?: exploratory analysis of EQ-5D-5L-derived utility values. Qual Life Res. 2018;27(3):717–24.

26. Pan CW, Zhang RY, Luo N, He JY, Liu RJ, Ying XH, et al. How the EQ-5D utilities are derived matters in Chinese diabetes patients: a comparison based on different EQ-5D scoring functions for China. *Qual Life Res.* 2020 Nov 1; **29**(11): 3087–94.

27. Yang F, Devlin N, Luo N. Cost-Utility Analysis Using EQ-5D-5L Data: Does How the Utilities Are Derived Matter? *Value Health*. 2019; **22**(1): 45–9.

28. Alava MH, Wailoo A, Grimm S, Pudney S, Gomes M, Sadique Z, et al. EQ-5D-5L versus EQ-5D-3L: The Impact on Cost Effectiveness in the United Kingdom. *Value in Health*. 2018 Jan 1; **21**(1): 49–56.

29. Ben Â, Finch AP, Dongen JM van, Wit M de, Dijk SEM van, Snoek FJ, et al. Comparing the EQ-5D-5L crosswalks and value sets for England, the Netherlands and Spain: Exploring their impact on cost-utility results. *Health Economics.* 2020; **29**(5): 640–51.

30. Kolovos S, Bosmans JE, van Dongen JM, van Esveld B, Magai D, van Straten A, et al. Utility scores for different health states related to depression: individual participant data analysis. *Qual Life Res*. 2017; **26**(7): 1649–58.

31. Maas ET, Juch JN, Groeneweg JG, Ostelo RW, Koes BW, Verhagen AP, et al. Cost-effectiveness of minimal interventional procedures for chronic mechanical low back pain: design of four randomised controlled trials with an economic evaluation. *BMC Musculoskeletal Disorders*. 2012 Dec 28; **13**(1): 260.

32. Mutubuki EN, van Helvoirt H, van Dongen JM, Vleggeert-Lankamp CLA, Huygen FJPM, van Tulder MW, et al. Cost-effectiveness of combination therapy (Mechanical Diagnosis and Treatment and Transforaminal Epidural Steroid Injections) among patients with an indication for a Lumbar Herniated Disc surgery: Protocol of a randomized controlled trial. *Physiotherapy Research International.* 2020; **25**(1): e1796.

33. Kloek CJ, Bossen D, Veenhof C, van Dongen JM, Dekker J, de Bakker DH. Effectiveness and cost-effectiveness of a blended exercise intervention for patients with hip and/or knee osteoarthritis: study protocol of a randomized controlled trial. *BMC Musculoskeletal Disorders*. 2014 Aug 8; **15**(1): 269.

34. Knoop J, Dekker J, van der Leeden M, de Rooij M, Peter WFH, van Bodegom-Vos L, et al. Stratified exercise therapy compared with usual care by physical therapists in patients with knee osteoarthritis: A randomized controlled trial protocol (OCTOPuS study). *Physiotherapy Research International.* 2020; **25**(2): e1819.

35. van Dongen JM, Persoon S, Jongeneel G, Bosmans JE, Kersten MJ, Brug J, et al. Long-term effectiveness and cost-effectiveness of an 18-week supervised exercise program in patients treated with autologous stem cell transplantation: results from the EXIST study. J Cancer Surviv [Internet]. 2019 Jul 8 [cited 2019 Jul 19]; Available from: https://doi.org/10.1007/s11764-019-00775-9

36. El Alili M, Schuurhuizen CSEW, Braamse AMJ, Beekman ATF, van der Linden MH, Konings IR, et al. Economic evaluation of a combined screening and stepped-care treatment program targeting psychological distress in patients with metastatic colorectal cancer: A cluster randomized controlled trial. *Palliat Med*. 2020 Jul 1; **34**(7): 934–45.

37. Carmody TJ, Rush AJ, Bernstein I, Warden D, Brannan S, Burnham D, et al. The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. *Eur Neuropsychopharmacol.* 2006 Dec; 1**6**(8): 601–11.

38. Trivedi MH, Rush AJ, Ibrahim HM, Carmody TJ, Biggs MM, Suppes T, et al. The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychol Med*. 2004 Jan; **34**(1): 73–82.

39. Snaith RP. The Hospital Anxiety And Depression Scale. *Health Qual Life Outcomes*. 2003 Aug 1; **1** :29.

40. Smarr KL, Keefer AL. Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9). *Arthritis Care & Research*. 2011; **63**(S11): S454–66.

41. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. *Annals of the Rheumatic Diseases*. 1978 Aug 1; **37**(4): 378–81.

42. Smets EMA, Garssen B, Bonke B, De Haes JCJM. The multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. *Journal of Psychosomatic Research.* 1995 Apr 1; **39**(3): 315–25.

43. Thong MSY, Mols F, van de Poll-Franse LV, Sprangers MAG, van der Rijt CCD, Barsevick AM, et al. Identifying the subtypes of cancer-related fatigue: results from the population-based PROFILES registry. *J Cancer Surviv.* 2018 Feb 1; **12**(1): 38–46.

44. Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol.* 1982 Dec; **5**(6): 649–55.

45. Parkin D, Zamora B, Feng Y, van Hout B, Devlin N. The EQ-5D simulation laboratory: a resource for testing 3L and 5L real and mapped value sets. *EuroQol group*; 2019.

46. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J: L. Erlbaum Associates; 1988. 567 p.

47. Goldfeld K. simstudy: Simulation of Study Data [Internet]. 2019 [cited 2019 May 24]. Available from: https://CRAN.R-project.org/package=simstudy

48. Morton F, Nijjar JS. eq5d: Methods for Analysing "EQ-5D" Data and Calculating "EQ-5D" Index Scores [Internet]. 2021 [cited 2021 Aug 18]. Available from: https://CRAN.R-project.org/package=eq5d

49. Walters SJ, Brazier JE. Comparison of the Minimally Important Difference for Two Health State Utility Measures: EQ-5D and SF-6D. *Quality of Life Research*. 2005; **14**(6): 1523–32.

50. Willan AR. Incremental net benefit in the analysis of economic data from clinical trials, with application to the CADET-Hp trial. *Eur J Gastroenterol Hepatol.* 2004 Jun; **16**(6): 543–9.

51. Black WC. The CE plane: a graphic representation of cost-effectiveness. *Med Decis Making*. 1990 Sep; **10**(3): 212–4.

52. Barber JA, Thompson SG. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine*. 2000; **19**(23): 3219–36.

53. Löthgren M, Zethraeus N. Definition, interpretation and calculation of cost-effectiveness acceptability curves. *Health Econ.* 2000 Oct; **9**(7): 623–30.

Appendix 1 | Cut-off points of severity levels

Health condition	Measure	Mild	Moderate	Severe
Depression	MADRS26, 33	0–26	27–34	35–60
	IDS-SR26, 34	0–38	39–48	49–84
	HADS-D26, 35	0–19	20–25	26–52
	BDI-II36	0–19	20–28	29–63
Low back pain and osteoarthritis	NRS37	0–3	4–6	7–10
Cancer	MSI38, 39	0–10	11–16	17–20
	ECOG40	0–1	2	3–4

MADRS: Montgomery–Åsberg Depression Rating Scale. IDS-SR: The Inventory of Depressive Symptomatology, Clinician Rating and Self-Report. HADS: Hospital Anxiety and Depression Scale. BDI-II: Beck Depression Inventory-II. NRS: Numeric Rating Scale. MSI: Multidimensional Fatigue Inventory. ECOG: Eastern Cooperative Oncology Group performance status.

Appendix 2 | R script – Data generation

```
# EQ-5D-3L Baseline Profile Generator for R
# Version 2.00 March 2019

# 1. load libraries
library(readr)
# 2. Set up a working diretory
setwd(“<<directory path>>”)

# 3. Load the function profGen.f to generate baseline profiles for the 3L version
profGen.f <<-function(cases){
 dim.name <- c(“Mobility”,”Self Care”,”Usual Act”,”Pain & Dis”,”Anx & Dep”,”Profile”)
 baseProf.m <<- matrix(0,nrow=cases, ncol=6)
 colnames(baseProf.m) <<- dim.name
 for (v1 in 1:cases) {
 prof=0
 for (v2 in 1:5) {
 P=runif(1)
 if(P <= prob.m[v2,1]) {
 baseProf.m[v1,v2] <<- 1
 }else{
 if(P <=(prob.m[v2,1]+prob.m[v2,2])) {
 baseProf.m[v1,v2] <<- 2
 }else { baseProf.m[v1,v2] <<- 3 }
 }
 prof=prof+(baseProf.m[v1,v2]*(10^(5-v2)))
 }
 baseProf.m[v1,6] <<- prof
 }
 cat(“\n”,”First 10 profiles”,”\n”)
 print(baseProf.m[1:10,1:6])
}
#********************************************************************************
## DISEASE (d) [i.e., DEPRESSION, LOW BACK PAIN, OSTEOARTHRITIS, CANCER]

## BASELINE PROFILE – MILD
## Import the disease baseline probability matrix control group
prob.m <- read.csv(“d_bpm_ml_t0.csv”, row.names=1)
profGen.f(150) # generate baseline profiles
write.csv(baseProf.m, “d_bp_ml_t0.csv”) # save baseline profiles
## Import the disease baseline probability matrix treatment group
prob.m <- read.csv(“d_bpm_ml_t1.csv”, row.names=1)
profGen.f(150)
write.csv(baseProf.m, “d_bp_ml_t1.csv”)

## BASELINE PROFILE – MODERATE
## Import the disease baseline probability matrix control group
prob.m <- read.csv(“d_bpm_mo_t0.csv”, row.names=1)
```

```
profGen.f(150)
write.csv(baseProf.m, “d_bp_mo_t0.csv”)
## Import the disease baseline probability matrix treatment group
prob.m <- read.csv(“d_bpm_mo_t1.csv”, row.names=1)
profGen.f(150)
write.csv(baseProf.m, “d_bp_mo_t1.csv”)

## BASELINE PROFILE – SEVERE
## Import the disease baseline probability matrix control group
prob.m <- read.csv(“d_bpm_se_t0.csv”, row.names=1)
profGen.f(150)
write.csv(baseProf.m, “d_bp_se_t0.csv”)
## Import the disease baseline probability matrix treatment group
prob.m <- read.csv(“d_bpm_se_t1.csv”, row.names=1)
profGen.f(150)
write.csv(baseProf.m, “d_bp_se_t1.csv”)

################################################################################
# Package: EQSimLab3L
# Title: EQ-5D-3L Simulation Laboratory
# Version: 0.0.0.9000

# 1. Load libraries
library(readr)
library(eq5d)
library(foreign)

# 2. Set up a working diretory
setwd(“<<directory path>>”)

# 3. generate follow-up profiles by treatment group
## TREATMENT GOUP
eq_make_profile_change()
# Do you want to import probabilities from a data file? (y = yes) > y
# File Name? > 0-tp-mld3l-t1.txt
# # Base profile data file name ? > d_bp_mlt1.txt
# Matched outcome profile data file name ? > d_fp_mlt1.txt

## CONTROL GROUP - the same transition probabilities as mild depression small effect size
eq_make_profile_change()
# Do you want to import probabilities from a data file? (y = yes) > y
# File Name? > 0-tp-mld3l-t0.txt
# Base profile data file name ? > d_bp_mlt0.txt
# Matched outcome profile data file name ? > d_fp_mlt0.txt

# 4. Calculate baseline and follow-up utilities
# EQ-5D-5L country-specific value sets must be one of: Canada, China, Denmark, Egypt, England,
# Ethiopia, France, Germany, HongKong, Hungary, Indonesia, Ireland, Japan, Malaysia, Netherlands,
# Peru_cTTO, Peru_DCE, Poland, Portugal, SouthKorea, Spain, Sweden, Taiwan, Thailand, Uruguay,
# USA, Vietnam
# 4.1 baseline utilities control group
`baseProf.m.t0` <- read.csv(“~/2020-EuroQol-call/EQ5D_laboratory/probabilities/d_bp_mlt0.txt”, sep=””)
names(baseProf.m.t0)[1] <- “id”
names(baseProf.m.t0)[2] <- “MO”
names(baseProf.m.t0)[3] <- “SC”
names(baseProf.m.t0)[4] <- “UA”
names(baseProf.m.t0)[5] <- “PD”
names(baseProf.m.t0)[6] <- “AD”
names(baseProf.m.t0)[7] <- “Profile.b”
baseProf.m.t0$trt <- 0
baseProf.m.t0$NL.utility.b <- eq5d(baseProf.m.t0, type=”TTO”, version=”3L”, country = “Netherlands”, ignore.incomplete = TRUE)
baseProf.m.t0$US.utility.b <- eq5d(baseProf.m.t0, type=”TTO”, version=”3L”, country = “USA”, ignore.incomplete = TRUE)
baseProf.m.t0$JP.utility.b <- eq5d(baseProf.m.t0, type=”TTO”, version=”3L”, country = “Japan”, ignore.incomplete = TRUE)

# 4.2 follow-up utilities control group
`outProf.m.t0` <- read.csv(“~/2020-EuroQol-call/EQ5D_laboratory/probabilities/d_fp_mlt0.txt”, sep=””)
names(outProf.m.t0)[1] <- “id”
names(outProf.m.t0)[2] <- “MO”
names(outProf.m.t0)[3] <- “SC”
names(outProf.m.t0)[4] <- “UA”
names(outProf.m.t0)[5] <- “PD”
names(outProf.m.t0)[6] <- “AD”
names(outProf.m.t0)[7] <- “Profile.f”
outProf.m.t0$trt <- 0
outProf.m.t0 <- data.frame(lapply(outProf.m.t0, function(x) as.numeric(as.character(x))))
outProf.m.t0$NL.utility.f <- eq5d(outProf.m.t0, type=”TTO”, version=”3L”, country = “Netherlands”, ignore.incomplete = TRUE)
outProf.m.t0$US.utility.f <- eq5d(outProf.m.t0, type=”TTO”, version=”3L”, country = “USA”, ignore.incomplete = TRUE)
outProf.m.t0$JP.utility.f <- eq5d(outProf.m.t0, type=”TTO”, version=”3L”, country = “Japan”, ignore.incomplete = TRUE)

# 4.3. Merge baseline and follow-up data control group
control <- merge(baseProf.m.t0, outProf.m.t0, by = “id”)

# 4.4 baseline utilities treatment group
`baseProf.m.t1` <- read.csv(“~/2020-EuroQol-call/EQ5D_laboratory/probabilities/dep_bp_mlt1.txt”, sep=””)
names(baseProf.m.t1)[1] <- “id”
names(baseProf.m.t1)[2] <- “MO”
names(baseProf.m.t1)[3] <- “SC”
names(baseProf.m.t1)[4] <- “UA”
names(baseProf.m.t1)[5] <- “PD”
names(baseProf.m.t1)[6] <- “AD”
names(baseProf.m.t1)[7] <- “Profile.b”
baseProf.m.t1$trt <- 1
baseProf.m.t1$NL.utility.b <- eq5d(baseProf.m.t1, type=”TTO”, version=”3L”, country = “Netherlands”, ignore.incomplete = TRUE)
baseProf.m.t1$US.utility.b <- eq5d(baseProf.m.t1, type=”TTO”, version=”3L”, country = “USA”, ignore.incomplete = TRUE)
baseProf.m.t1$JP.utility.b <- eq5d(baseProf.m.t1, type=”TTO”, version=”3L”, country = “Japan”, ignore.incomplete = TRUE)

# 4.5 follow-up utilities treatment group
`outProf.m.t1` <- read.csv(“~/2020-EuroQol-call/EQ5D_laboratory/probabilities/dep_fp_mlt1.txt”, sep=””)
names(outProf.m.t1)[1] <- “id”
names(outProf.m.t1)[2] <- “MO”
names(outProf.m.t1)[3] <- “SC”
names(outProf.m.t1)[4] <- “UA”
names(outProf.m.t1)[5] <- “PD”
names(outProf.m.t1)[6] <- “AD”
names(outProf.m.t1)[7] <- “Profile.f”
outProf.m.t1$trt <- 0
outProf.m.t1 <- data.frame(lapply(outProf.m.t1, function(x) as.numeric(as.character(x))))
outProf.m.t1$NL.utility.f <- eq5d(outProf.m.t1, type=”TTO”, version=”3L”, country = “Netherlands”, ignore.incomplete = TRUE)
outProf.m.t1$US.utility.f <- eq5d(outProf.m.t1, type=”TTO”, version=”3L”, country = “USA”, ignore.incomplete = TRUE)
outProf.m.t1$JP.utility.f <- eq5d(outProf.m.t1, type=”TTO”, version=”3L”, country = “Japan”, ignore.incomplete = TRUE)

# 4.6. Merge baseline and follow-up data treatment group
treatment <- merge(baseProf.m.t1, outProf.m.t1, by = “id”)

# 5. Merge treatment and control
small.eff.size <- rbind(treatment, control)

# 6. Calculate cohen’s d = mean difference between groups/ sd
small.eff.size$NL.QALY <- (0.5*(small.eff.size$NL.utility.b + small.eff.size$NL.utility.f))
small.eff.size$US.QALY <- (0.5*(small.eff.size$US.utility.b + small.eff.size$US.utility.f))
small.eff.size$JP.QALY <- (0.5*(small.eff.size$US.utility.b + small.eff.size$JP.utility.f))

# 7. Prepare data to save in .dta
colnames(small.eff.size)[colnames(small.eff.size)==”trt.x”] <- “trt”
colnames(small.eff.size)[colnames(small.eff.size)==”MO.x”] <- “MO_BASELINE”
colnames(small.eff.size)[colnames(small.eff.size)==”SC.x”] <- “SC_BASELINE”
colnames(small.eff.size)[colnames(small.eff.size)==”UA.x”] <- “UA_BASELINE”
colnames(small.eff.size)[colnames(small.eff.size)==”PD.x”] <- “PD_BASELINE”
colnames(small.eff.size)[colnames(small.eff.size)==”AD.x”] <- “AD_BASELINE”
colnames(small.eff.size)[colnames(small.eff.size)==”MO.y”] <- “MO_T1”
colnames(small.eff.size)[colnames(small.eff.size)==”SC.y”] <- “SC_T1”
colnames(small.eff.size)[colnames(small.eff.size)==”UA.y”] <- “UA_T1”
colnames(small.eff.size)[colnames(small.eff.size)==”PD.y”] <- “PD_T1”
colnames(small.eff.size)[colnames(small.eff.size)==”AD.y”] <- “AD_T1”
small.eff.size$trt.y <- NULL

# 8. Check effect size NL, US, JP
NL.lm <-lm(NL.QALY ~ trt, data = small.eff.size)
summary(NL.lm)
sd.NL <- sd(small.eff.size$NL.QALY)
```

```
cohen.d.NL <- NL.lm[[“coefficients”]][[“trt”]]/sd.NL
cohen.d.NL

US.lm <-lm(US.QALY ~ trt, data = small.eff.size)
summary(US.lm)
sd.US <- sd(small.eff.size$US.QALY)
cohen.d.US <-US.lm[[“coefficients”]][[“trt”]]/sd.US
cohen.d.US

JP.lm <-lm(JP.QALY ~ trt, data = small.eff.size)
summary(JP.lm)
sd.JP <- sd(small.eff.size$JP.QALY)
cohen.d.JP <- JP.lm[[“coefficients”]][[“trt”]]/sd.JP
cohen.d.JP

small.eff.size <<- cbind(Case =c(1:nrow(small.eff.size)),small.eff.size)
small.eff.size$id <- NULL
colnames(small.eff.size)[colnames(small.eff.size)==”Case”] <- “id”

write.dta(small.eff.size, file = “<<directory path>>/ml-small-effsize.dta”)

################################################################################
# Generate age, gender, and costs
# 1. Load libraries
library(haven)
library(simstudy)
library(foreign)

# 2. Import EQ-5D dataset and prepare to merge with simulated dataset
setwd(“<<directory path>>”)
dataset <- read_dta(“ml-small-effsize.dta”) # replace the name of EQ-5D dataset here

# 3. Generate baseline characteristics
def <- defData(varname = “age”, dist=”uniformInt”, formula=”25;75”, id=”id”)
def <- defData(def, varname = “gender”, formula = 0.19, dist = “binary”, id=”id”)
simulatie <- genData(300, def)


# 4. Merge baseline characteristics and EQ-5D dataset
simulatie <- merge(simulatie,dataset, by =”id”)

# 5. Generate correlated costs and QALYs
def1 <- defDataAdd(varname = “costs”, formula = “2000 + 250*trt”, variance = 1, dist = “gamma”)
simulatie <- addColumns(def1, simulatie)
simulatie <- addCorFlex(simulatie, def1, rho = 0.75, corstr = “cs”)

# 6. Check correlation
correlation <- simulatie[,cor(NL_QALY, i.costs)]
correlation
```

```
# 7. Save data in .dta
write.dta(simulatie, file = “<<directory path>>/ml-small-effsize.dta”)
################################################################################
# Package: EQSimLab5L
# Title: EQ-5D-5L Simulation Laboratory
# Version: 0.0.0.9000

# 1. Load libraries
library(EQSimLab5L)
library(readr)
library(dplyr)
library(eq5d)
library(foreign)

# 2. Set up a working diretory
setwd(“<<directory path>>”)

# 3. Generate baseline profiles per treatment group
## TREATMENT GOUP
eq_make_profile_data(150)
#Do you want to import probabilities? (y = yes) > n
# These are the probabilities that you have specified:
# Level 1 Level 2 Level 3 Level 4 Level 5
# Mobility 0.00 0.25 0.50 0.25 0
# Self Care 0.50 0.25 0.25 0.00 0
# Usual Activities 0. 00 0.25 0.25 0.50 0
# Pain & Discomfort 0.00 0.00 0.50 0.50 0
# Anxiety & Depression 0.25 0.50 0.00 0.25 0
# Save probability matrix? (y = yes) > y
# Save File Name? > d_bpm_ml5l_t1.txt
# Do you want to randomise by Profile or Dimension? (Choose p for Profile) > p
# Save File Name? > d_bp_ml5l_t1.txt

## CONTROL GROUP
eq_make_profile_data(150)
#Do you want to import probabilities? (y = yes) > n
# These are the probabilities that you have specified:
# Level 1 Level 2 Level 3 Level 4 Level 5
# Mobility 0.00 0.00 0.75 0.25 0
# Self Care 0.25 0.50 0.25 0.00 0
# Usual Activities 0.25 0.00 0.25 0.50 0
# Pain & Discomfort 0.00 0.00 0.75 0.25 0
# Anxiety & Depression 0.25 0.25 0.50 0.00 0
# Save probability matrix? (y = yes) > y
# Save File Name? > d_bpm_ml5l_t0.txt
# Do you want to randomise by Profile or Dimension? (Choose p for Profile) > p
# Save File Name? > d_bp_ml5l_t0.txt

# 4. Generate follow-up profiles per treatment group
## TREATMENT GOUP
eq_make_profile_change()
# Do you want to import probabilities from a data file? (y = yes) > y
# File Name? > d_tp_ml5l_t1.txt
# Base profile data file name ? > d_bp_ml5l_t1.txt
# Matched outcome profile data file name ? > d_fp_ml5l_t1.txt

## CONTROL GROUP
eq_make_profile_change()
# Do you want to import probabilities from a data file? (y = yes) > y
# File Name? > d_tp_ml5l_t0.txt
# File name ? > d_bp_ml5l_t0.txt
# Matched outcome profile data file name ? > d_fp_ml5l_t0.txt

# 5. Calculate baseline and follow-up utilities
# 5.1 baseline utilities control group
`baseProf.m.t0` <- read.csv(“<<directory path>>/d_fp_ml5l_t0.txt”, sep=“”)
names(baseProf.m.t0)[1] <- “id”
names(baseProf.m.t0)[2] <- “MO”
names(baseProf.m.t0)[3] <- “SC”
names(baseProf.m.t0)[4] <- “UA”
names(baseProf.m.t0)[5] <- “PD”
names(baseProf.m.t0)[6] <- “AD”
names(baseProf.m.t0)[7] <- “Profile.b”
baseProf.m.t0$trt <- 0
baseProf.m.t0$NL.utility.b <- eq5d(baseProf.m.t0, type=”VT”, version=”5L”, country = “Netherlands”, ignore.incomplete
= TRUE)
baseProf.m.t0$US.utility.b <- eq5d(baseProf.m.t0, type=”VT”, version=”5L”, country = “USA”, ignore.incomplete = TRUE)
baseProf.m.t0$JP.utility.b <- eq5d(baseProf.m.t0, type=”VT”, version=”5L”, country = “Japan”, ignore.incomplete =
TRUE)

# 5.2 follow-up utilities control group
`outProf.m.t0` <- read.csv(“<<directory path>>/d_fp_ml5l_t0.txt”, sep=“”)
names(outProf.m.t0)[1] <- “id”
names(outProf.m.t0)[2] <- “MO”
names(outProf.m.t0)[3] <- “SC”
names(outProf.m.t0)[4] <- “UA”
names(outProf.m.t0)[5] <- “PD”
names(outProf.m.t0)[6] <- “AD”
names(outProf.m.t0)[7] <- “Profile.f”
outProf.m.t0$trt <- 0
outProf.m.t0 <- data.frame(lapply(outProf.m.t0, function(x) as.numeric(as.character(x))))
outProf.m.t0$NL.utility.f <- eq5d(outProf.m.t0, type=”VT”, version=”5L”, country = “Netherlands”, ignore.incomplete
= TRUE)
outProf.m.t0$US.utility.f <- eq5d(outProf.m.t0, type=”VT”, version=”5L”, country = “USA”, ignore.incomplete = TRUE)
outProf.m.t0$JP.utility.f <- eq5d(outProf.m.t0, type=”VT”, version=”5L”, country = “Japan”, ignore.incomplete = TRUE)
```

```
# 5.3. Merge baseline and follow-up data control group
control <- merge(baseProf.m.t0, outProf.m.t0, by = “id”)

# 5.4 baseline utilities treatment group
`baseProf.m.t1` <- read.csv(“<<directory path>>/d_fp_ml5l_t1.txt”, sep=””)
names(baseProf.m.t1)[1] <- “id”
names(baseProf.m.t1)[2] <- “MO”
names(baseProf.m.t1)[3] <- “SC”
names(baseProf.m.t1)[4] <- “UA”
names(baseProf.m.t1)[5] <- “PD”
names(baseProf.m.t1)[6] <- “AD”
names(baseProf.m.t1)[7] <- “Profile.b”
baseProf.m.t1$trt <- 1
baseProf.m.t1$NL.utility.b <- eq5d(baseProf.m.t1, type=”VT”, version=”5L”, country = “Netherlands”, ignore.incomplete = TRUE)
baseProf.m.t1$US.utility.b <- eq5d(baseProf.m.t1, type=”VT”, version=”5L”, country = “USA”, ignore.incomplete = TRUE)
baseProf.m.t1$JP.utility.b <- eq5d(baseProf.m.t1, type=”VT”, version=”5L”, country = “Japan”, ignore.incomplete = TRUE)

# 5.5 follow-up utilities treatment group
`outProf.m.t1` <- read.csv(“<<directory path>>/d_fp_ml5l_t1.txt”, sep=””)
names(outProf.m.t1)[1] <- “id”
names(outProf.m.t1)[2] <- “MO”
names(outProf.m.t1)[3] <- “SC”
names(outProf.m.t1)[4] <- “UA”
names(outProf.m.t1)[5] <- “PD”
names(outProf.m.t1)[6] <- “AD”
names(outProf.m.t1)[7] <- “Profile.f”
outProf.m.t1$trt <- 0
outProf.m.t1 <- data.frame(lapply(outProf.m.t1, function(x) as.numeric(as.character(x))))

outProf.m.t1$NL.utility.f <- eq5d(outProf.m.t1, type=”VT”, version=”5L”, country = “Netherlands”, ignore.incomplete = TRUE)
outProf.m.t1$US.utility.f <- eq5d(outProf.m.t1, type=”VT”, version=”5L”, country = “USA”, ignore.incomplete = TRUE)
outProf.m.t1$JP.utility.f <- eq5d(outProf.m.t1, type=”VT”, version=”5L”, country = “Japan”, ignore.incomplete = TRUE)

# 5.6. Merge baseline and follow-up data treatment group
treatment <- merge(baseProf.m.t1, outProf.m.t1, by = “id”)

# 6. Merge treatment and control
small.eff.size <- rbind(treatment, control)
```

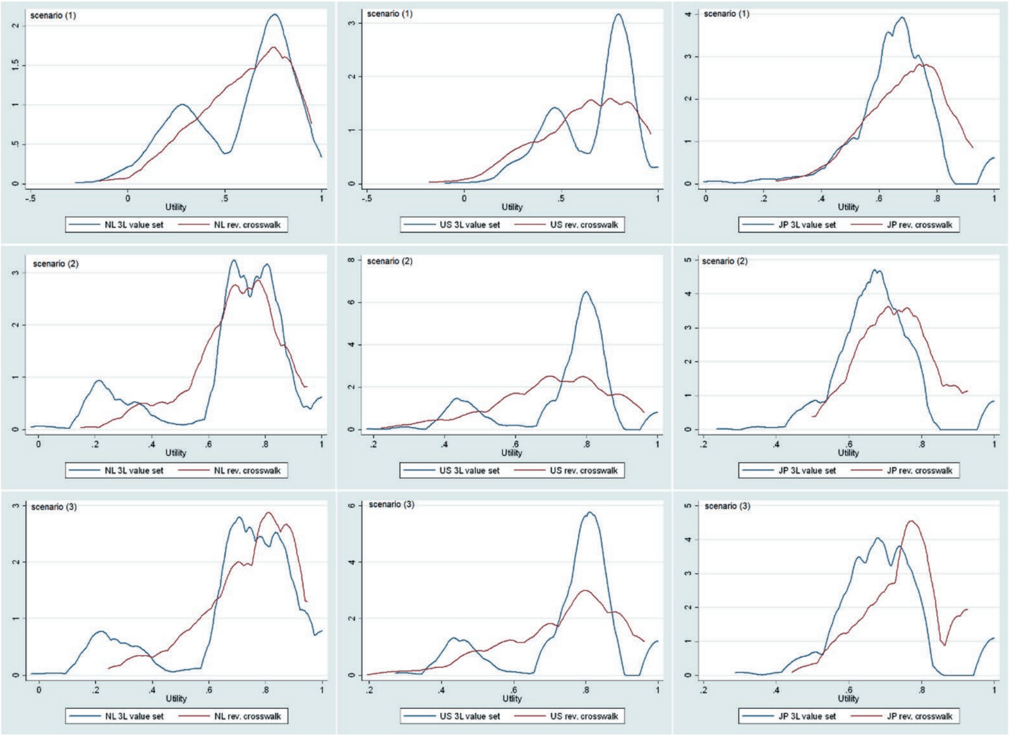
```
# 7. Calculate QALY
small.eff.size$NL.QALY <- (0.5*(small.eff.size$NL.utility.b + small.eff.size$NL.utility.f))
small.eff.size$JP.QALY <- (0.5*(small.eff.size$US.utility.b + small.eff.size$JP.utility.f))
small.eff.size$US.QALY <- (0.5*(small.eff.size$US.utility.b + small.eff.size$US.utility.f))

# 8. Check effect size (cohen’s d = mean difference bewteen groups/ sd)
colnames(small.eff.size)[colnames(small.eff.size)==”trt.x”] <- “trt”

NL.lm <-lm(NL.QALY ~ trt, data = small.eff.size)
summary(NL.lm)
sd.NL <- sd(small.eff.size$NL.QALY)
cohen.d.NL <- NL.lm[[“coefficients”]][[“trt”]]/sd.NL
cohen.d.NL

US.lm <-lm(US.QALY ~ trt, data = small.eff.size)
summary(US.lm)
sd.US <- sd(small.eff.size$US.QALY)
cohen.d.US <-US.lm[[“coefficients”]][[“trt”]]/sd.US
cohen.d.US

JP.lm <-lm(JP.QALY ~ trt, data = small.eff.size)
summary(JP.lm)
sd.JP <- sd(small.eff.size$JP.QALY)
cohen.d.JP <- JP.lm[[“coefficients”]][[“trt”]]/sd.JP
cohen.d.JP

# 9. Prepare data to save in .dta
colnames(small.eff.size)[colnames(small.eff.size)==”MO.x”] <- “MO_BASELINE”
colnames(small.eff.size)[colnames(small.eff.size)==”SC.x”] <- “SC_BASELINE”
colnames(small.eff.size)[colnames(small.eff.size)==”UA.x”] <- “UA_BASELINE”
colnames(small.eff.size)[colnames(small.eff.size)==”PD.x”] <- “PD_BASELINE”
colnames(small.eff.size)[colnames(small.eff.size)==”AD.x”] <- “AD_BASELINE”
colnames(small.eff.size)[colnames(small.eff.size)==”MO.y”] <- “MO_T1”
colnames(small.eff.size)[colnames(small.eff.size)==”SC.y”] <- “SC_T1”
colnames(small.eff.size)[colnames(small.eff.size)==”UA.y”] <- “UA_T1”
colnames(small.eff.size)[colnames(small.eff.size)==”PD.y”] <- “PD_T1”
colnames(small.eff.size)[colnames(small.eff.size)==”AD.y”] <- “AD_T1”
small.eff.size$trt.y <- NULL

write.dta(small.eff.size, file = “<<directory path>>/ml-small-effsize.dta”)
```

```
#########################################################################################
# Generate age, gender, and costs
# 1. Load libraries
library(haven)
library(simstudy)
library(foreign)

# 2. Import EQ-5D dataset and prepare to merge with simulated dataset
dataset <- read_dta("<<directory path>>/ml-small-effsize.dta") # replace the name of EQ-5D dataset here
dataset <- as.data.frame(dataset)
dataset <- remove_label(dataset)

# 3. Generate baseline characteristics
def <- defData(varname = "age", dist="uniformInt", formula="25;75", id="id")
def <- defData(def, varname = "gender", formula = 0.19, dist = "binary", id="id")
simulatie <- genData(300, def)

# 4. Merge baseline characteristics and EQ-5D dataset
simulatie <- cbind(simulatie, dataset)
names(simulatie)[4] <- "id.x"
simulatie$id.x <- NULL

# 5. Generate correlated costs and QALYs
def1 <- defDataAdd(varname = "costs", formula = "2000 + 250*trt", variance = 1, dist = "gamma")
simulatie <- addColumns(def1, simulatie)

# 6. Check correlation
correlation <- simulatie[,cor(NL_QALY, costs)]
correlation

# 7. Save data in .dta
write.dta(simulatie, file = "<<directory path>>/ml-small-effsize.dta")
```

Appendix 3 | Kernel density histograms

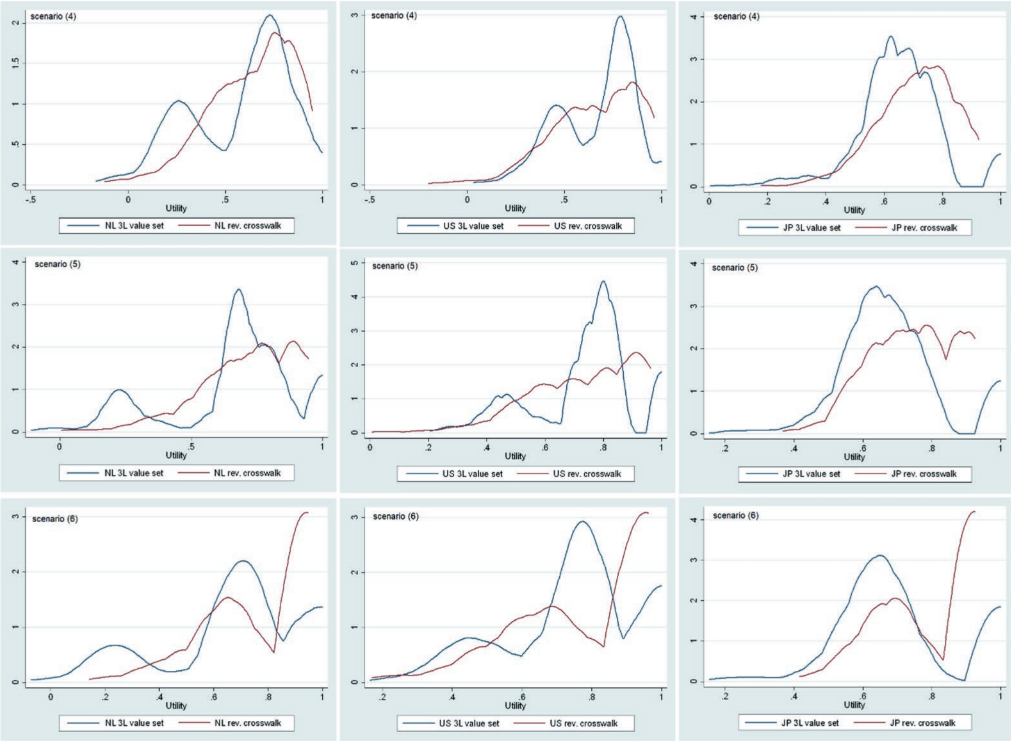
Kernel density histograms comparing utility distributions of the 3L value sets and 3L to 5L crosswalks.

MILD DEPRESSION



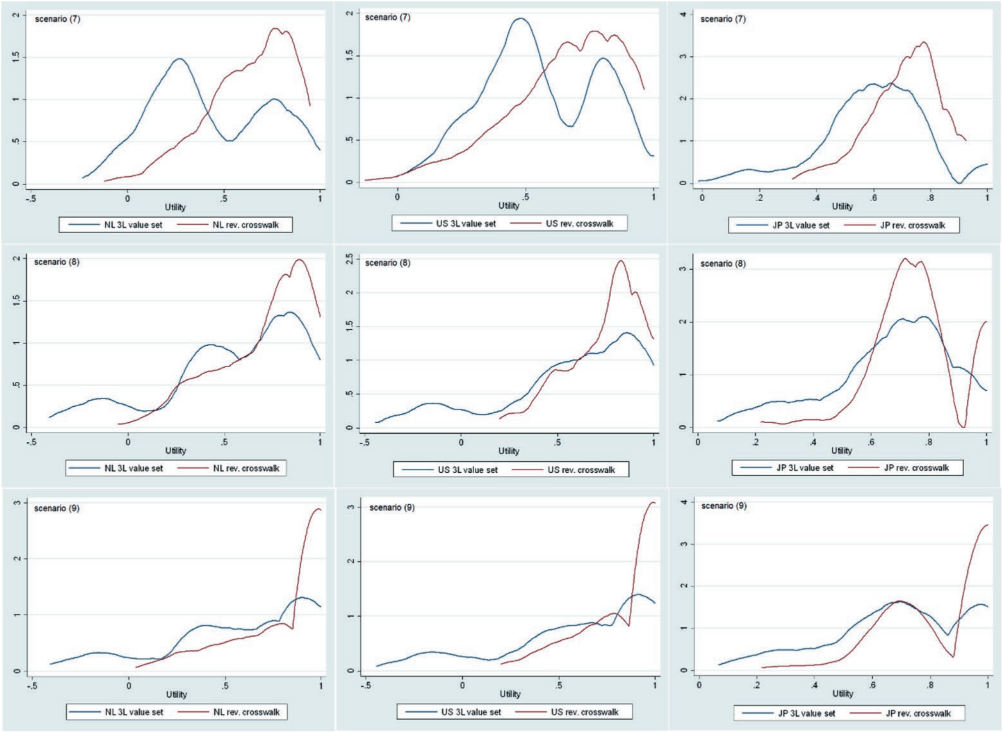
Scenario (1): mild depression and small treatment effect size.
Scenario (2): mild depression and medium treatment effect size.
Scenario (3): mild depression and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 3L to 5L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

MODERATE DEPRESSION



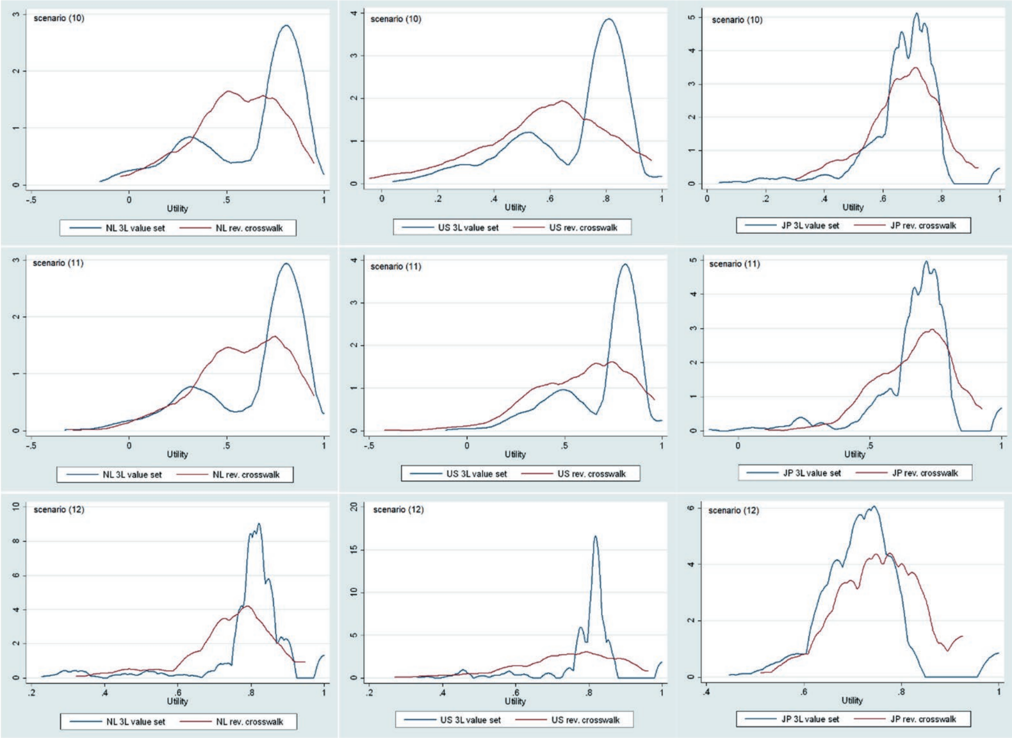
Scenario (4): moderate depression and small treatment effect size.
Scenario (5): moderate depression and medium treatment effect size.
Scenario (6): moderate depression and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

SEVERE DEPRESSION



Scenario (7): severe depression and small treatment effect size.
Scenario (8): severe depression and medium treatment effect size.
Scenario (9): severe depression and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

MILD LOW BACK PAIN



Scenario (10): mild low back pain and small treatment effect size.

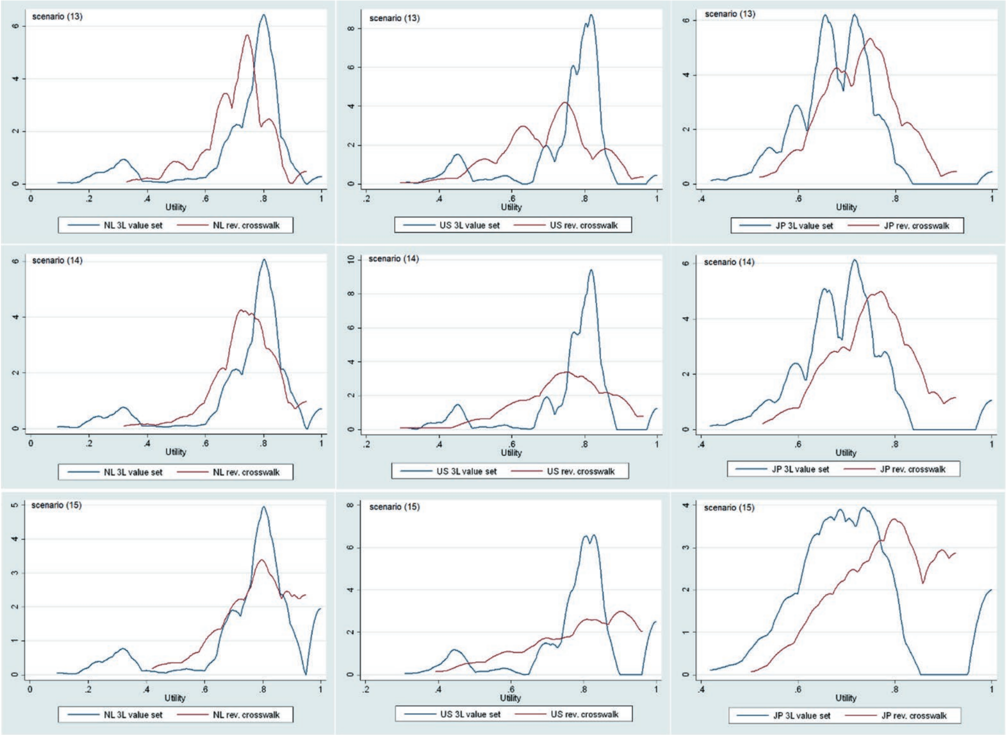
Scenario (11): mild low back pain and medium treatment effect size.

Scenario (12): mild low back pain and large treatment effect size.

3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk

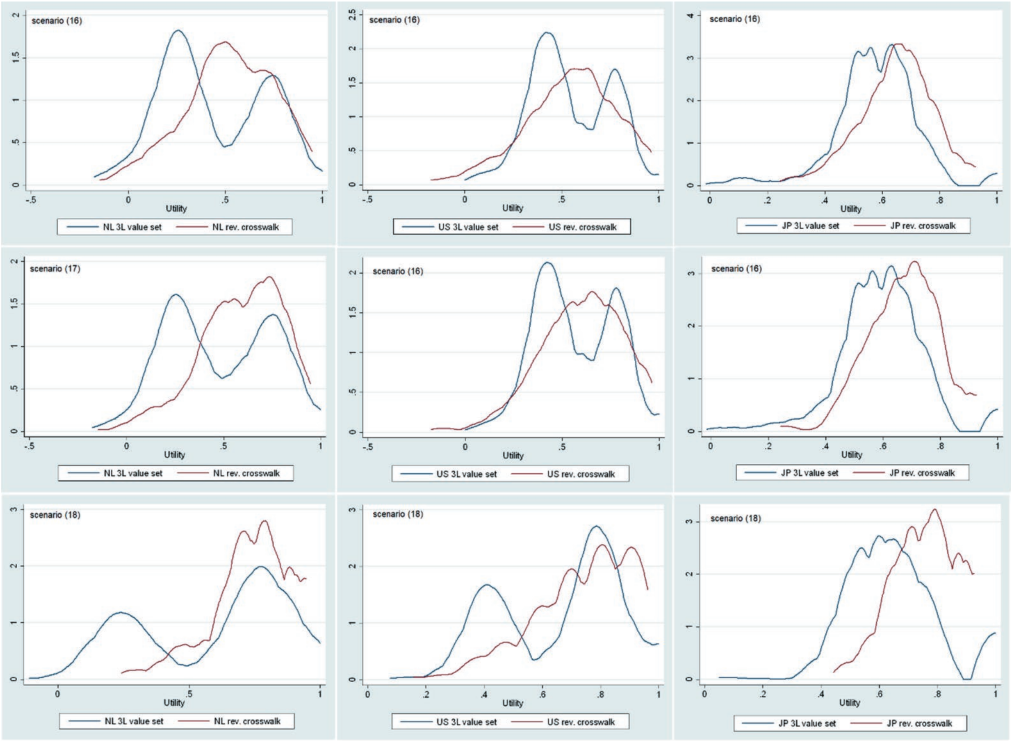
NL: the Netherlands. US: the United States. JP: Japan.

MODERATE LOW BACK PAIN



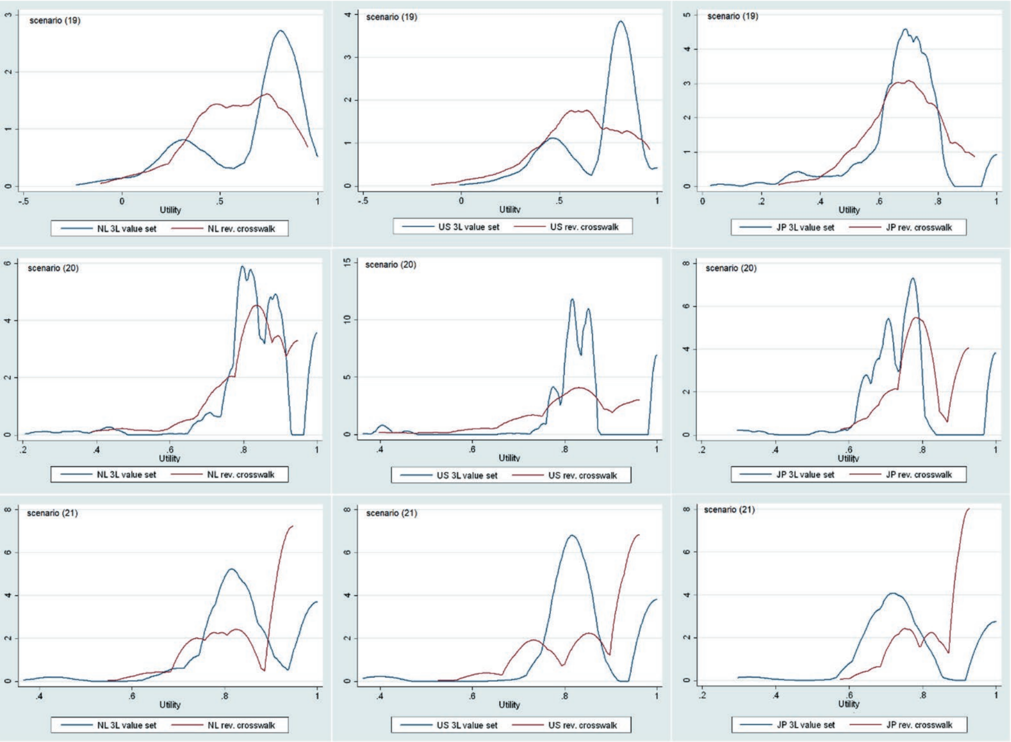
Scenario (13): moderate low back pain and small treatment effect size.
Scenario (14): moderate low back pain and medium treatment effect size.
Scenario (15): moderate low back pain and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

SEVERE LOW BACK PAIN



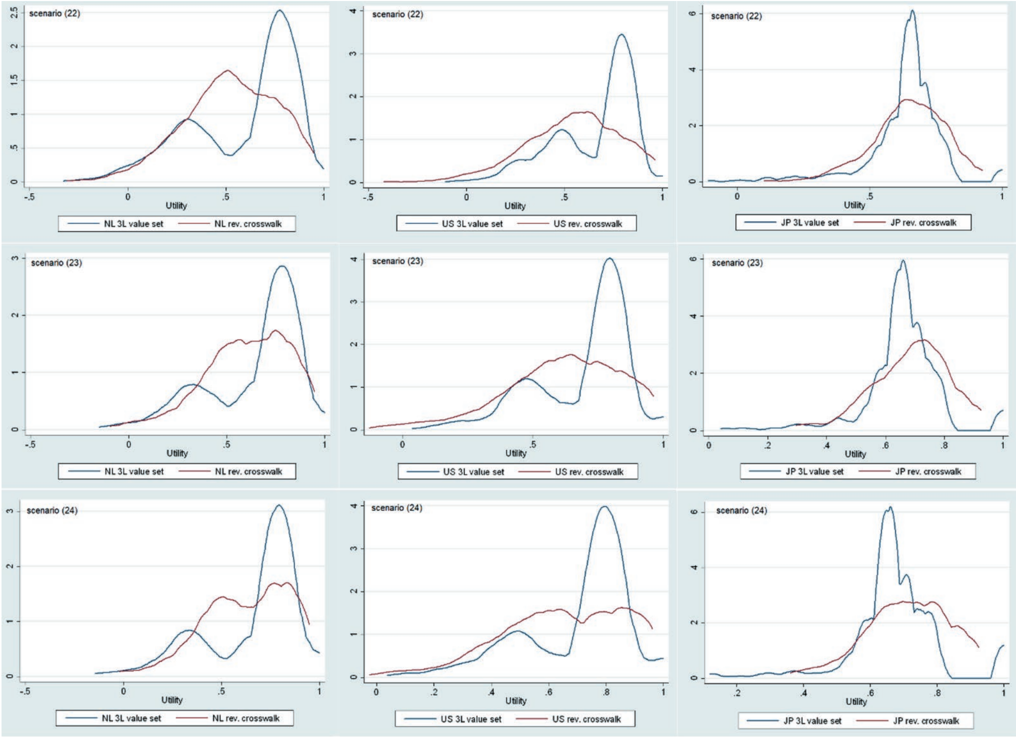
Scenario (16): severe low back pain and small treatment effect size.
Scenario (17): severe low back pain and medium treatment effect size.
Scenario (18): severe low back pain and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

MILD OSTEOARTHRITIS



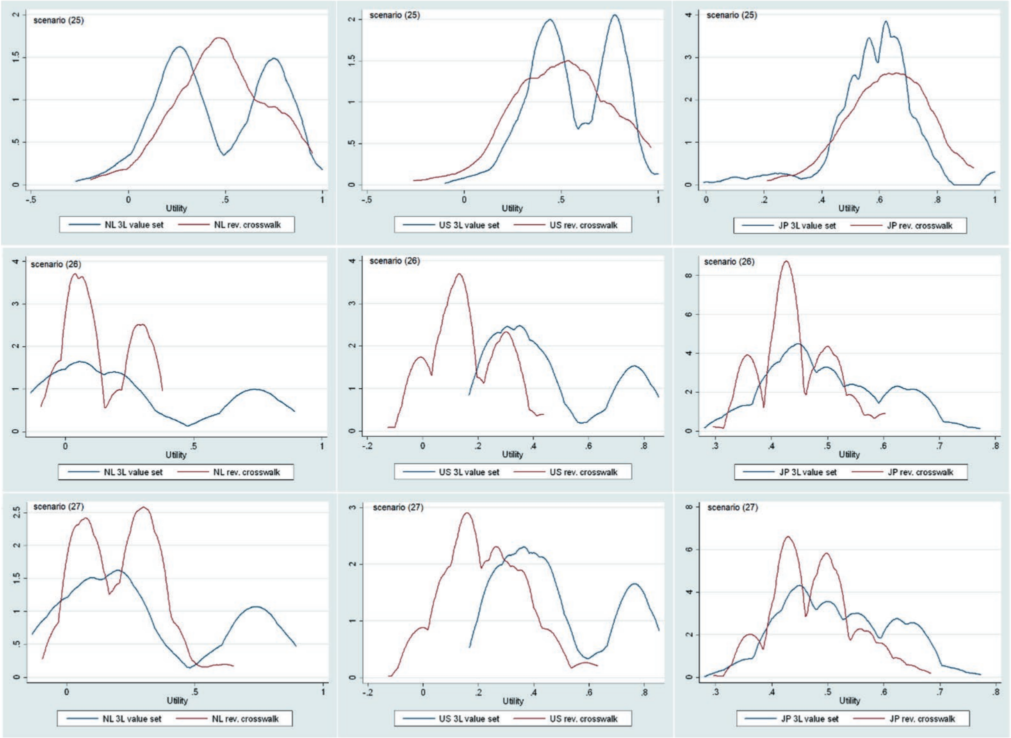
Scenario (19): mild osteoarthritis and small treatment effect size.
Scenario (20): mild osteoarthritis and medium treatment effect size.
Scenario (21): mild osteoarthritis pain and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

MODERATE OSTEOARTHRITIS



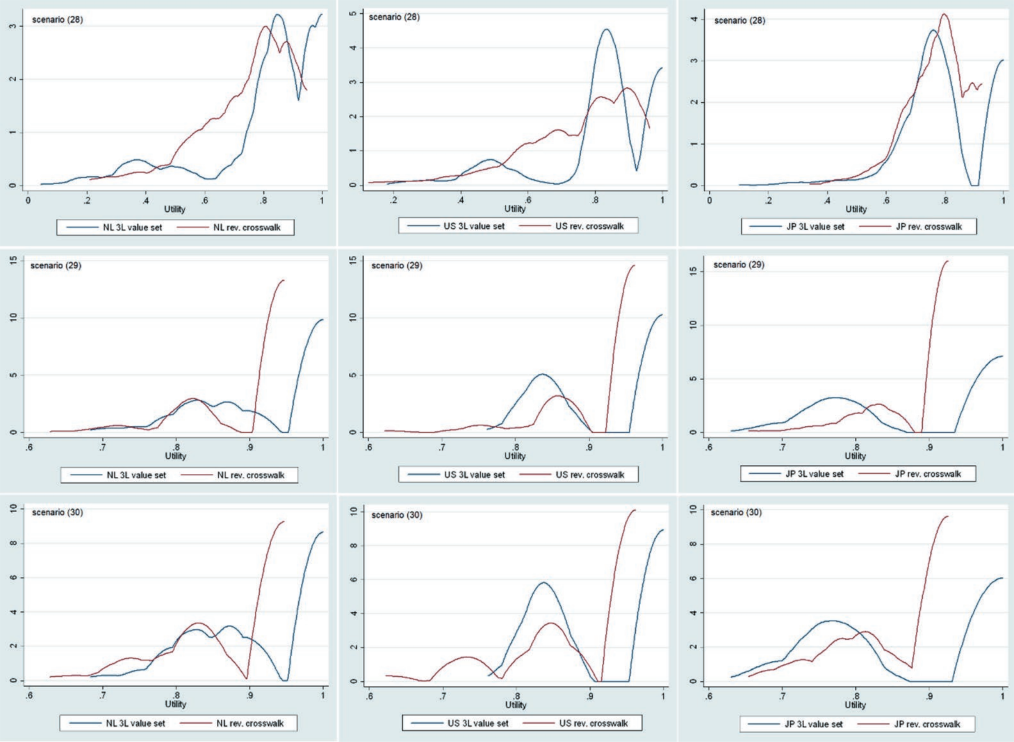
Scenario (22): moderate osteoarthritis and small treatment effect size.
Scenario (23): moderate osteoarthritis and medium treatment effect size.
Scenario (24): moderate osteoarthritis pain and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

SEVERE OSTEOARTHRITIS



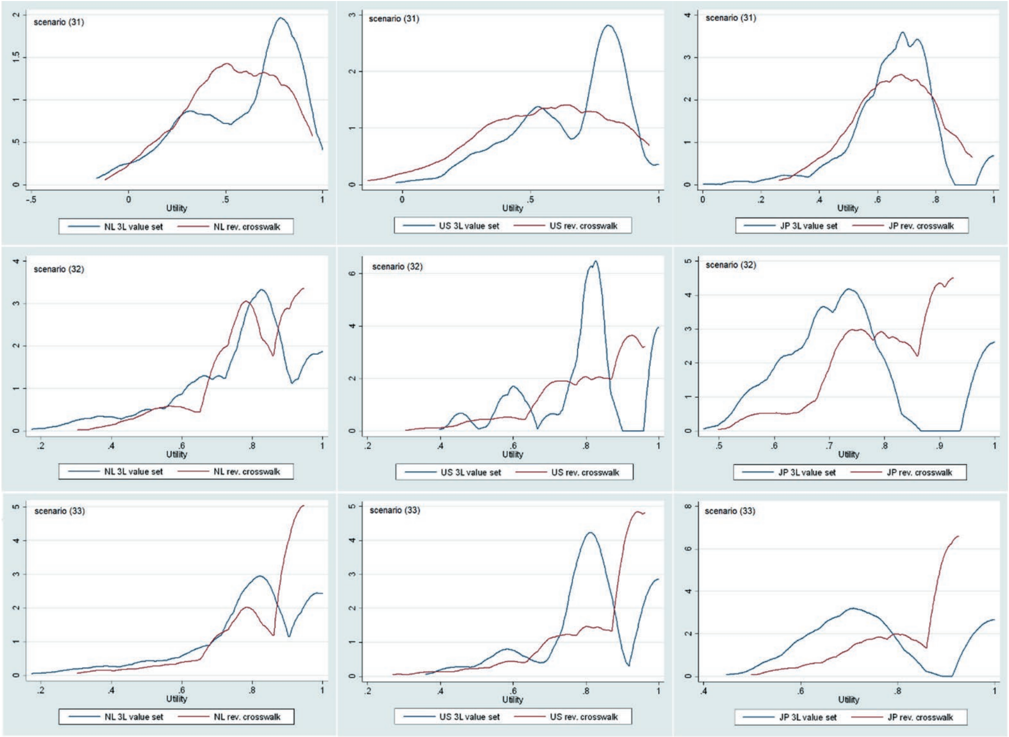
Scenario (25): severe osteoarthritis and small treatment effect size.
Scenario (26): severe osteoarthritis and medium treatment effect size.
Scenario (27): severe osteoarthritis pain and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

MILD CANCER



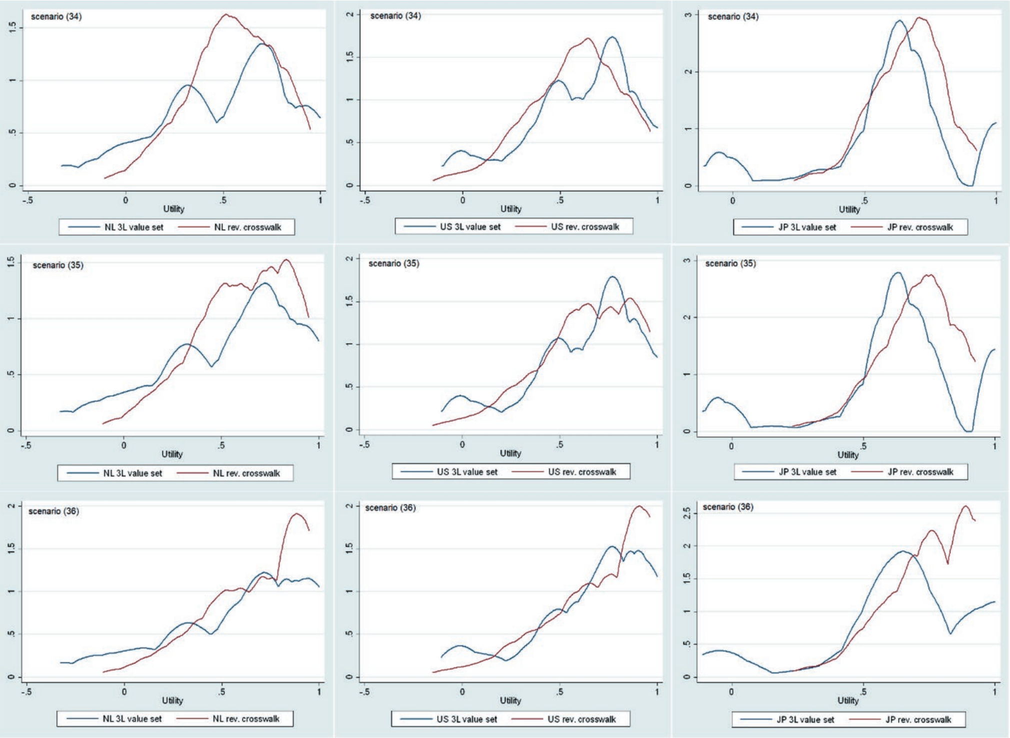
Scenario (28): mild cancer and small treatment effect size.
Scenario (29): mild cancer and medium treatment effect size.
Scenario (30): mild cancer and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

MODERATE CANCER



Scenario (31): moderate cancer and small treatment effect size.
Scenario (32): moderate cancer and medium treatment effect size.
Scenario (33): moderate cancer and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

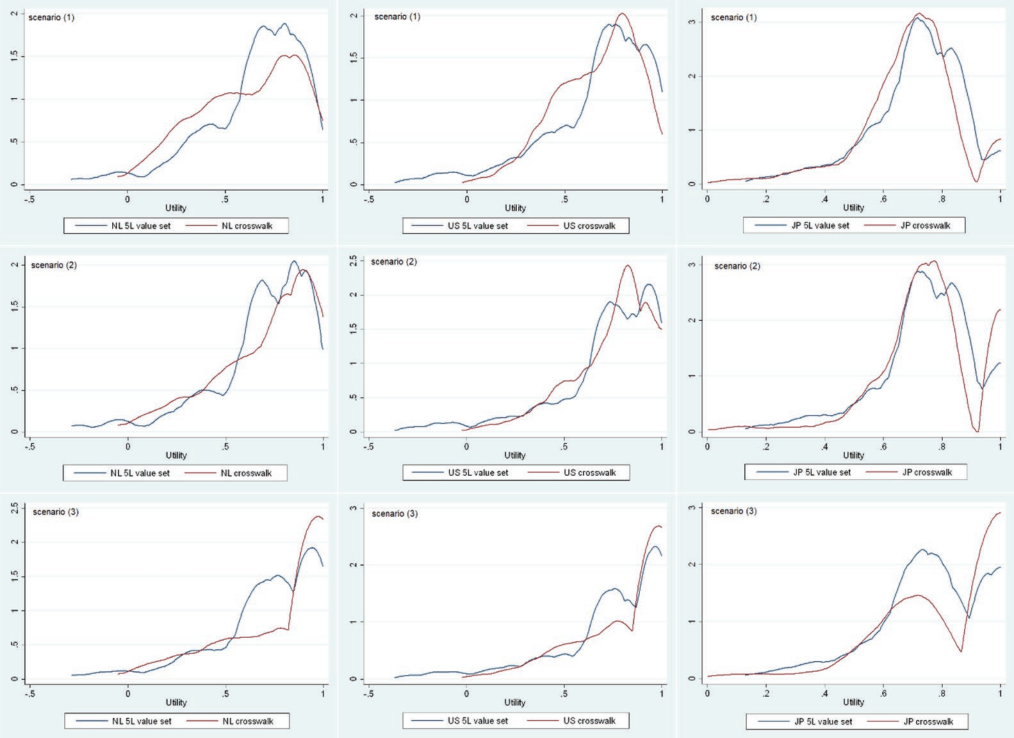
SEVERE CANCER



Scenario (34): severe cancer and small treatment effect size.
Scenario (35): severe cancer and medium treatment effect size.
Scenario (36): severe cancer and large treatment effect size.
3L value set: EQ-5D-3L value set. rev. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

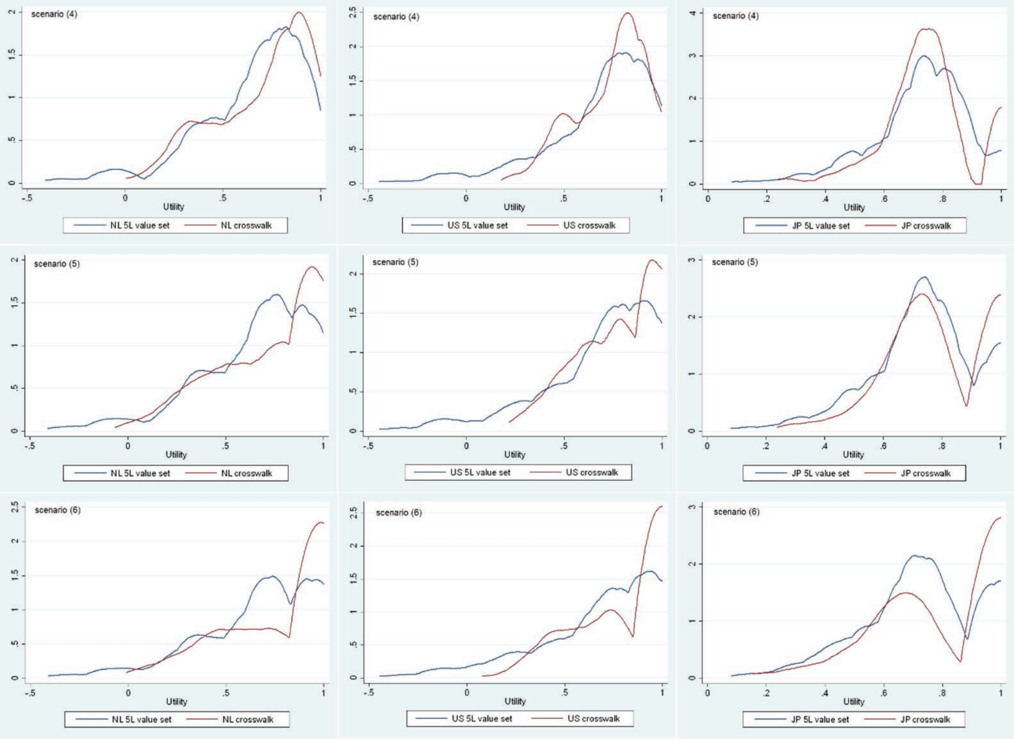
Kernel density histograms comparing utility distributions of the 5L value sets and 5L to 3L crosswalks.

MILD DEPRESSION



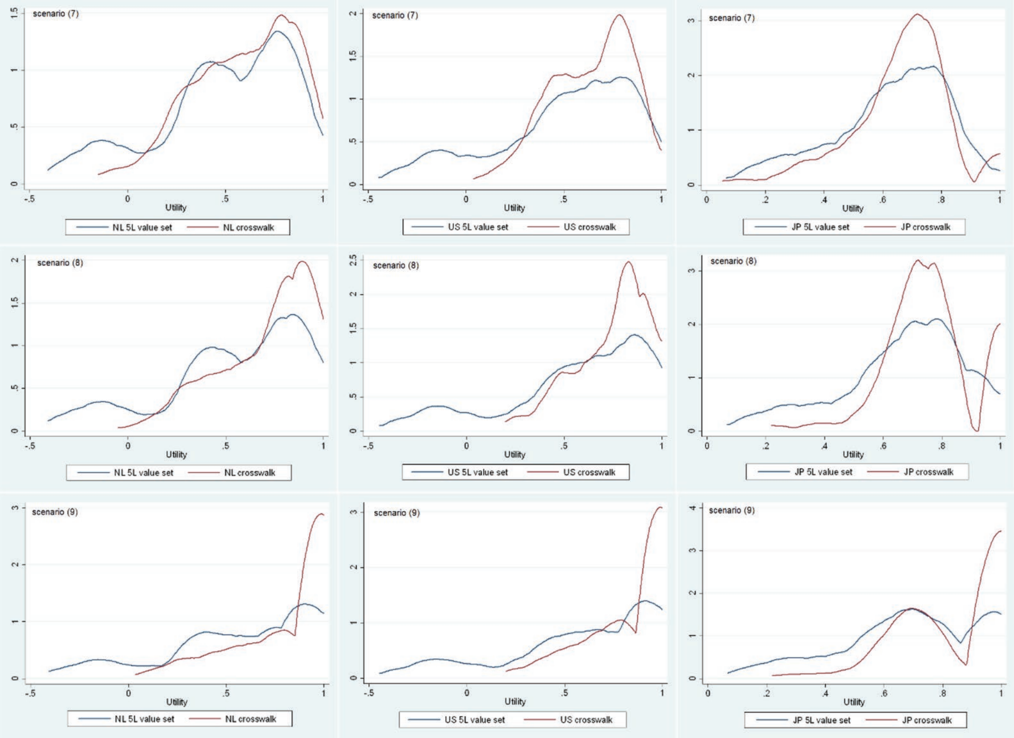
Scenario (1): mild depression and small treatment effect size.
Scenario (2): mild depression and medium treatment effect size.
Scenario (3): mild depression and large treatment effect size.
5L value set: EQ-5D-5L value set. crosswalk: 5L to 3L crosswalk
NL: the Netherlands. US: the United States. JP: Japan.

MODERATE DEPRESSION



Scenario (4): moderate depression and small treatment effect size.
Scenario (5): moderate depression and medium treatment effect size.
Scenario (6): moderate depression and large treatment effect size.
5L value set: EQ-5D-5L value set.
NL: the Netherlands. US: the United States. JP: Japan.

SEVERE DEPRESSION



Scenario (7): severe depression and small treatment effect size.

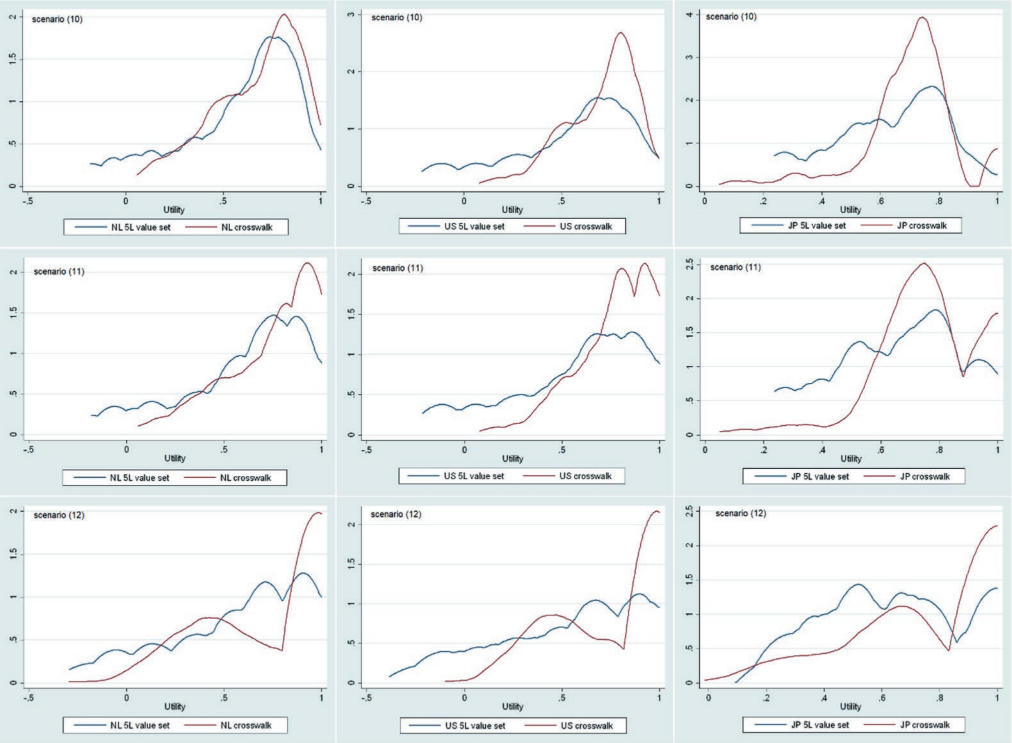
Scenario (8): severe depression and medium treatment effect size.

Scenario (9): severe depression and large treatment effect size.

5L value set: EQ-5D-5L value set.

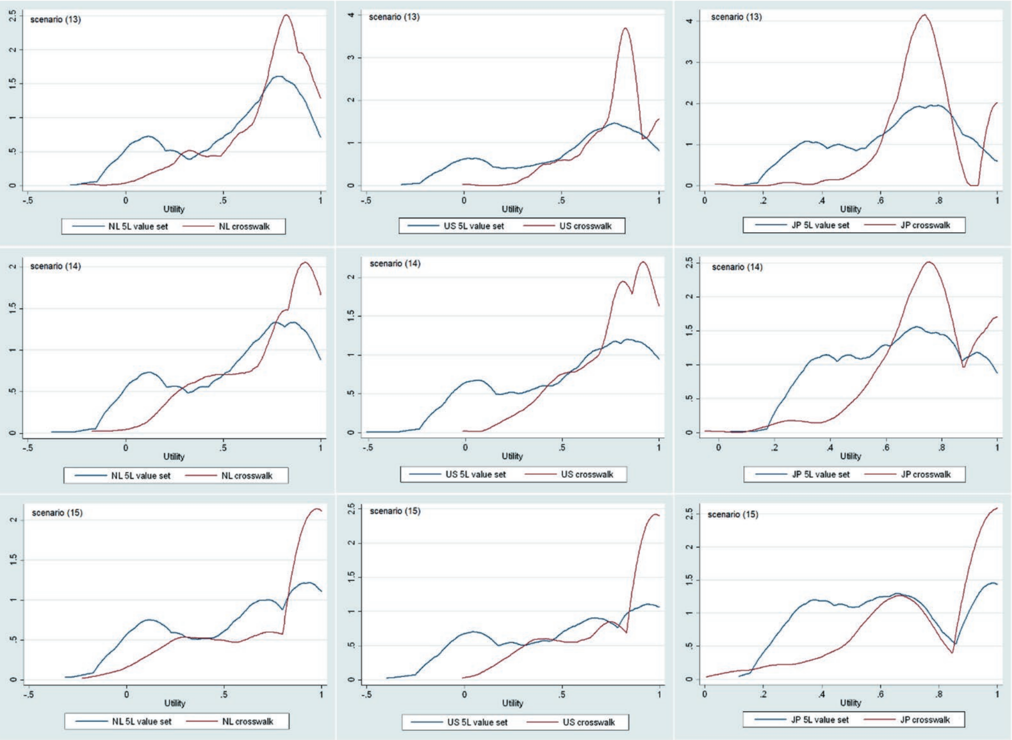
NL: the Netherlands. US: the United States. JP: Japan.

MILD LOW BACK PAIN



Scenario (10): mild low back pain and small treatment effect size.
Scenario (11): mild low back pain and medium treatment effect size.
Scenario (12): mild low back pain and large treatment effect size.
5L value set: EQ-5D-5L value set.
NL: the Netherlands. US: the United States. JP: Japan.

MODERATE LOW BACK PAIN



Scenario (13): moderate low back pain and small treatment effect size.

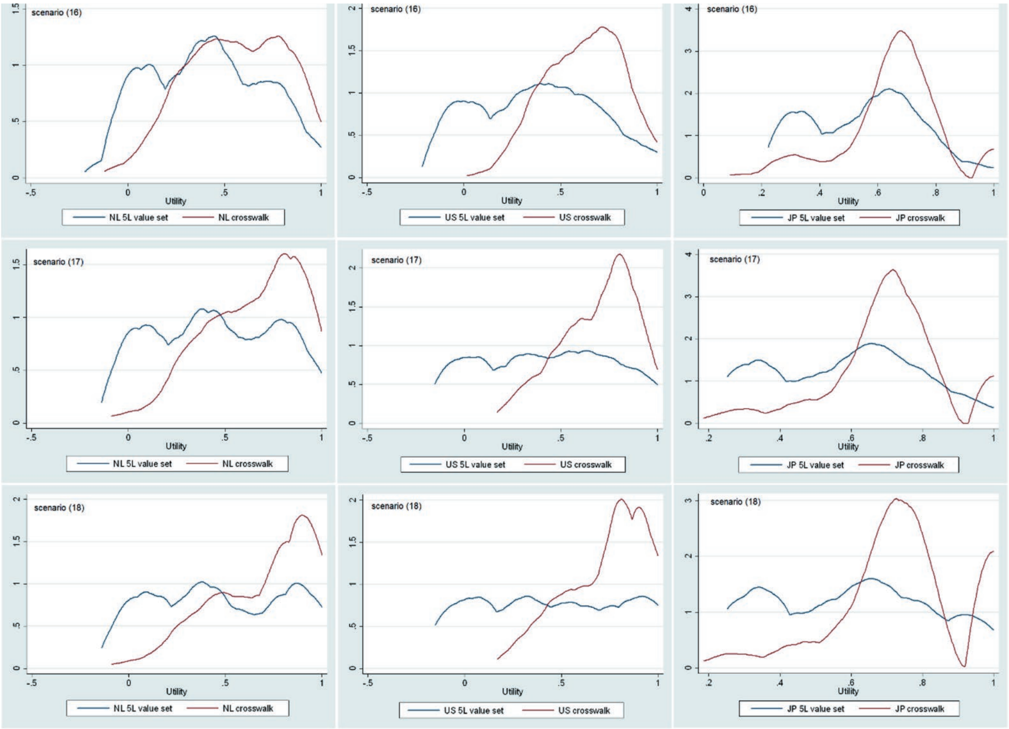
Scenario (14): moderate low back pain and medium treatment effect size.

Scenario (15): moderate low back pain and treatment large effect size.

5L value set: EQ-5D-5L value set.

NL: the Netherlands. US: the United States. JP: Japan.

SEVERE LOW BACK PAIN



Scenario (16): severe low back pain and small treatment effect size.

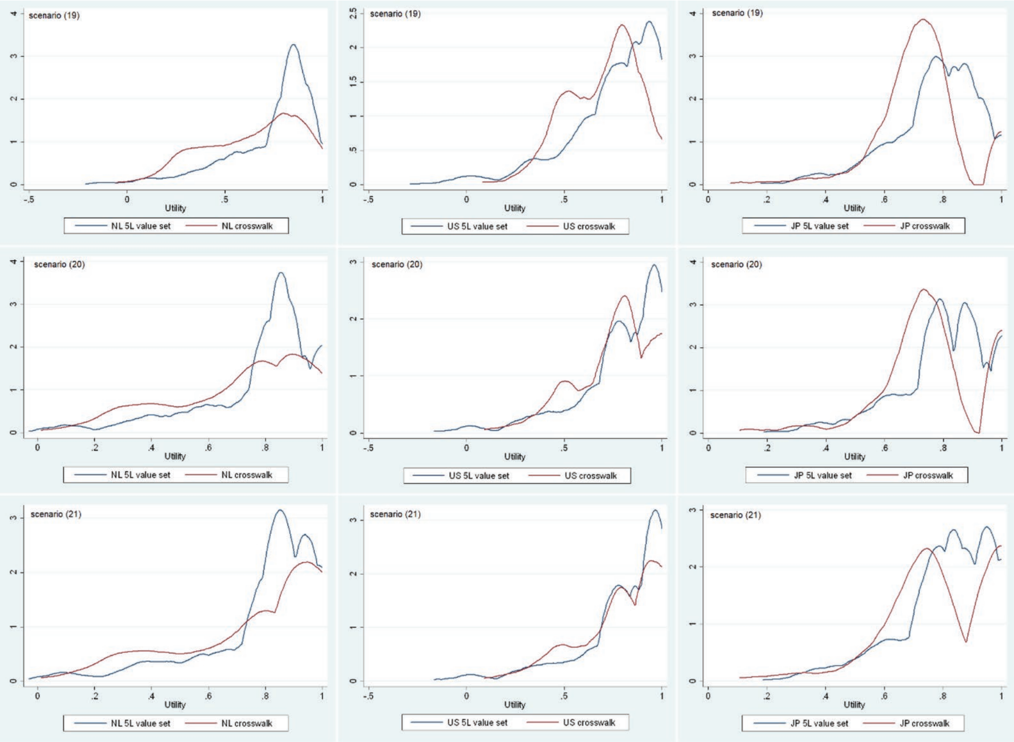
Scenario (17): severe low back pain and medium treatment effect size.

Scenario (18): severe low back pain and large treatment effect size.

5L value set: EQ-5D-5L value set.

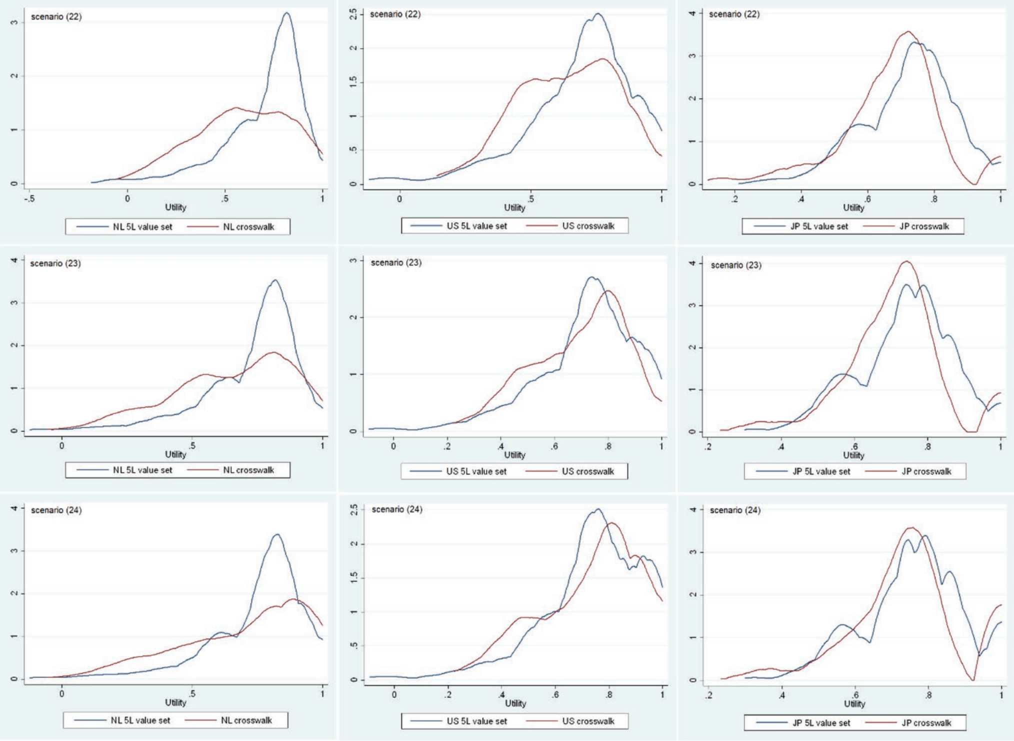
NL: the Netherlands. US: the United States. JP: Japan.

MILD OSTEOARTHRITIS



Scenario (19): mild osteoarthritis and small treatment effect size.
Scenario (20): mild osteoarthritis and medium treatment effect size.
Scenario (21): mild osteoarthritis pain and large treatment effect size.
5L value set: EQ-5D-5L value set.
NL: the Netherlands. US: the United States. JP: Japan.

MODERATE OSTEOARTHRITIS



Scenario (22): moderate osteoarthritis and small treatment effect size.

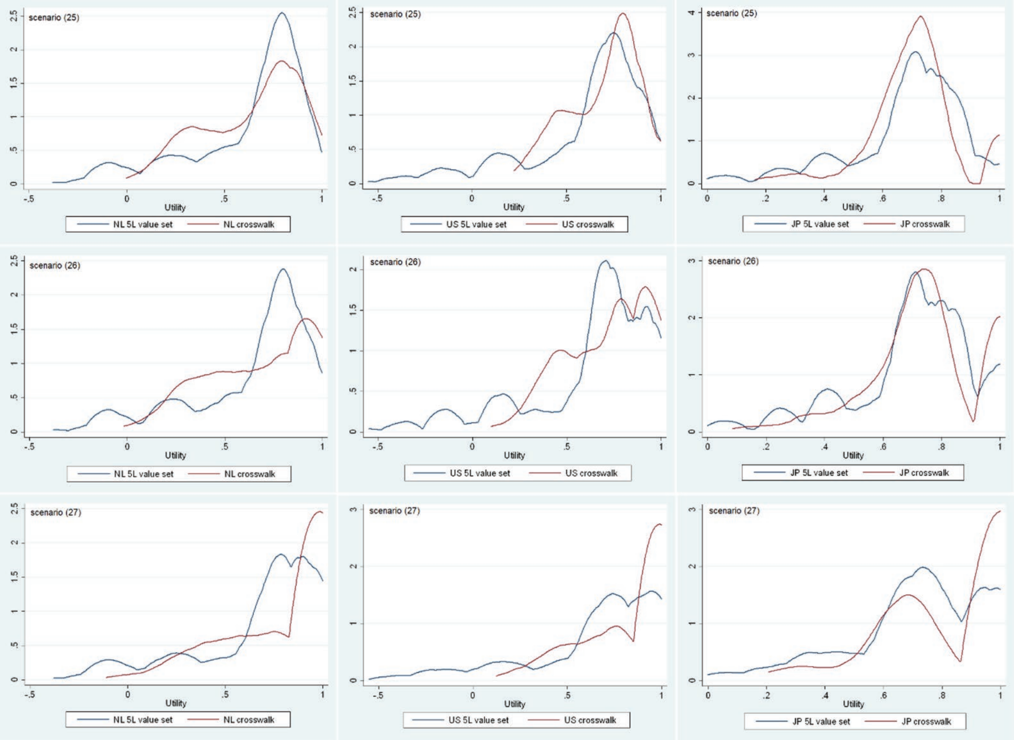
Scenario (23): moderate osteoarthritis and medium treatment effect size.

Scenario (24): moderate osteoarthritis pain and large treatment effect size.

5L value set: EQ-5D-5L value set.

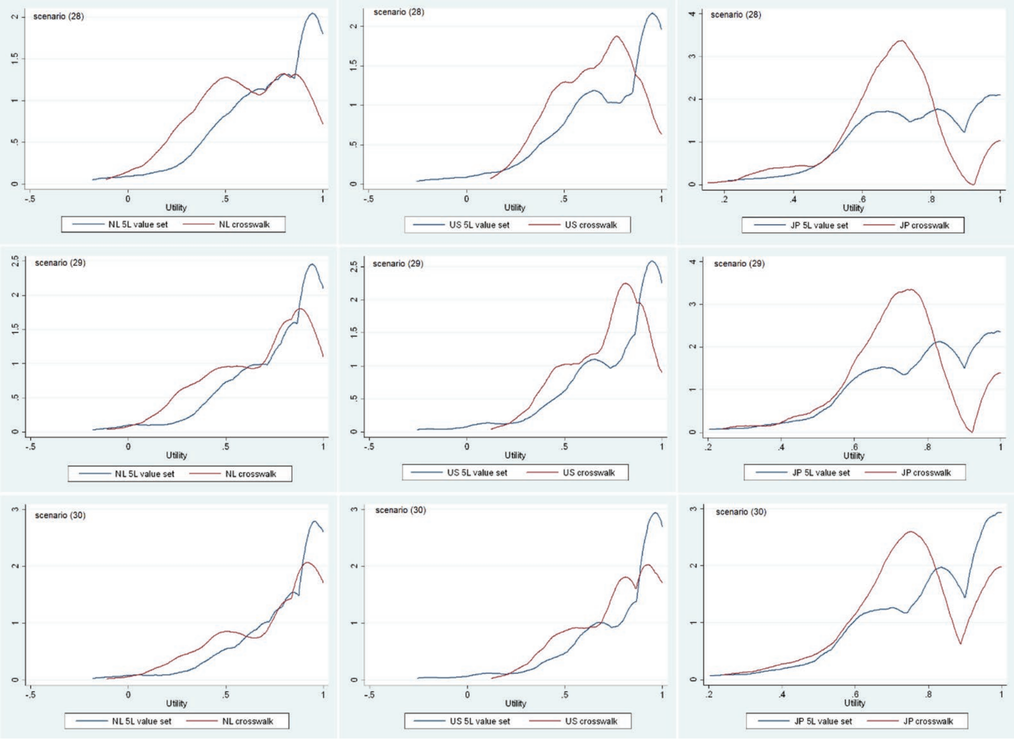
NL: the Netherlands. US: the United States. JP: Japan.

SEVERE OSTEOARTHRITIS



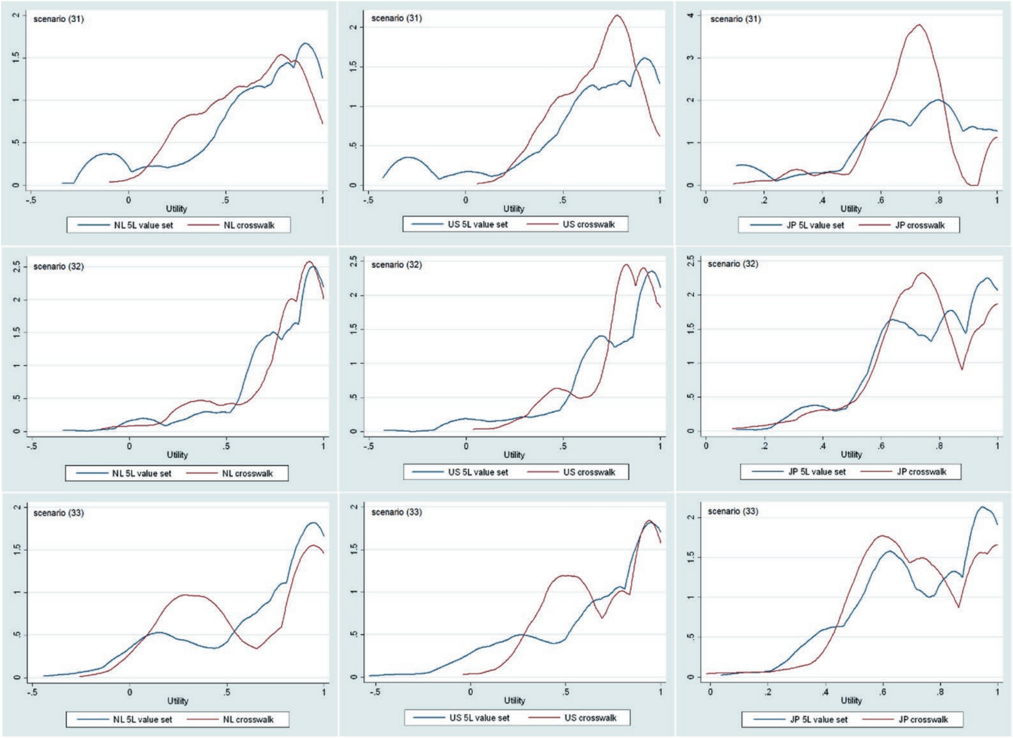
Scenario (25): severe osteoarthritis and small treatment effect size.
Scenario (26): severe osteoarthritis and medium treatment effect size.
Scenario (27): severe osteoarthritis pain and large treatment effect size.
5L value set: EQ-5D-5L value set.
NL: the Netherlands. US: the United States. JP: Japan.

MILD CANCER



Scenario (28): mild cancer and small treatment effect size.
Scenario (29): mild cancer and medium treatment effect size.
Scenario (30): mild cancer and large treatment effect size.
5L value set: EQ-5D-5L value set.
NL: the Netherlands. US: the United States. JP: Japan.

MODERATE CANCER



Scenario (31): moderate cancer and small treatment effect size.

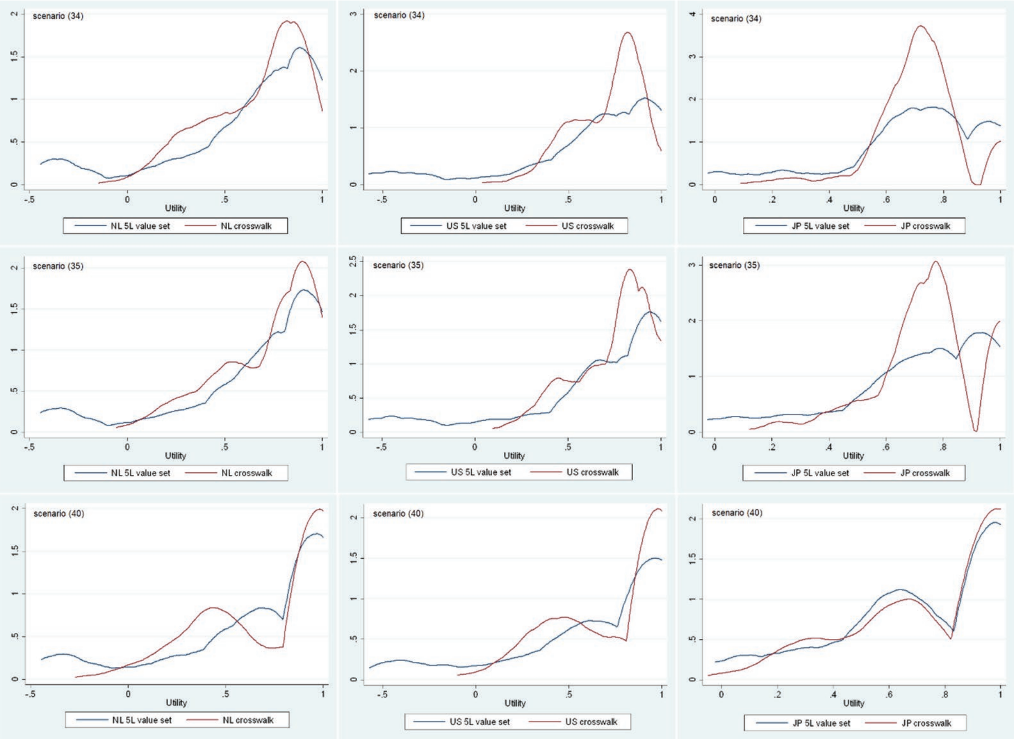
Scenario (32): moderate cancer and medium treatment effect size.

Scenario (33): moderate cancer and large treatment effect size.

5L value set: EQ-5D-5L value set.

NL: the Netherlands. US: the United States. JP: Japan.

SEVERE CANCER



Scenario (34): severe cancer and small treatment effect size.
Scenario (35): severe cancer and medium treatment effect size.
Scenario (36): severe cancer and large treatment effect size.
5L value set: EQ-5D-5L value set.
NL: the Netherlands. US: the United States. JP: Japan.

Appendix 4 | Results of cost-utility analyses.

Supplementary Table 4.1 | Cost-utility analysis results for 3L value set and 3L to 5L crosswalk per country

Country	Scoring method	Scenario	Patient population	Effect size	**Incremental QALYs** IQ (95% CI)	**Incremental costs** € (95% CI)	**ICER €/point**	**Distribution CE-plane (%)**				**Probability of cost-effectiveness**			
								NE	**SE**	**SW**	**NW**	**p_{CE}(0)**	**p_{CE}(20000)**	**p_{CE}(30000)**	**p_{CE}(50000)**
NL	3L value set	(1)	**Mild depression**	Small	0.05 (0.01; 0.09)	214 (-266; 654)	4656	81%	17%	0%	2%	0.173	0.911	0.949	0.968
	3L to 5L crosswalk				0.03 (-0.004; 0.06)	214 (-266; 654)	7284	80%	17%	1%	2%	0.173	0.813	0.883	0.933
US	3L value set				0.02 (-0.01; 0.06)	214 (-266; 654)	8781	76%	17%	1%	6%	0.173	0.731	**0.814**	**0.866**
	3L to 5L crosswalk				0.02 (-0.02; 0.05)	214 (-266; 654)	13971	68%	15%	2%	15%	0.173	0.583	**0.667**	**0.734**
JP	3L value set				0.02 (-0.01; 0.05)	214 (-266; 654)	11063	75%	17%	0%	8%	0.173	0.661	0.757	**0.830**
	3L to 5L crosswalk				0.0002 (-0.02; 0.02)	214 (-266; 654)	855681	40%	10%	7%	42%	0.173	0.242	0.290	**0.358**
NL	3L value set	(2)	**Mild depression**	Medium	0.08 (0.05; 0.12)	-27 (-619; 458)	-327	47%	53%	0%	0%	.526	1	1	1
	3L to 5L crosswalk				0.07 (0.04; 0.09)	-27 (-619; 458)	-407	47%	53%	0%	0%	.526	1	1	1
US	3L value set				0.05 (0.03; 0.08)	-27 (-619; 458)	-503	47%	53%	0%	0%	.526	.997	.998	.999
	3L to 5L crosswalk				0.06 (0.03; 0.09)	-27 (-619; 458)	-459	47%	53%	0%	0%	.526	.999	.999	1
JP	3L value set				0.06 (0.03; 0.08)	-27 (-619; 458)	-481	47%	53%	0%	0%	.526	.999	1	1
	3L to 5L crosswalk				0.04 (0.02; 0.05)	-27 (-619; 458)	-737	47%	53%	0%	0%	.526	.988	.998	1
NL	3L value set	(3)	**Mild depression**	Large	0.14 (0.10; 0.17)	-18 (-478; 442)	-139	47%	53%	0%	0%	0.530	1	1	1
	3L to 5L crosswalk				0.11 (0.08; 0.13)	-18 (-478; 442)	-178	47%	53%	0%	0%	0.530	1	1	1
US	3L value set				0.09 (0.07; 0.12)	-18 (-478; 442)	-203	47%	53%	0%	0%	0.530	1	1	1
	3L to 5L crosswalk				0.09 (0.07; 0.12)	-18 (-478; 442)	-205	47%	53%	0%	0%	0.530	1	1	1
JP	3L value set				0.09 (0.07; 0.12)	-18 (-478; 442)	-203	47%	53%	0%	0%	0.530	1	1	1
	3L to 5L crosswalk				0.06 (0.04; 0.07)	-18 (-478; 442)	-330	47%	53%	0%	0%	0.530	1	1	1

Country	Scoring method	Scenario	Patient population	Effect size	**Incremental QALYs** IQ (95% CI)	**Incremental costs** € (95% CI)	**ICER €/point**	**Distribution CE-plane (%)**				**Probability of cost-effectiveness**			
								NE	**SE**	**SW**	**NW**	**p_{CE}(0)**	**p_{CE}(20000)**	**p_{CE}(30000)**	**p_{CE}(50000)**
NL	3L value set	(4)	**Moderate depression**	Small	0.04 (0.0002; 0.08)	165 (-391; 651)	4129	73%	24%	0%	3%	0.246	0.893	0.930	0.960
	3L to 5L crosswalk				0.04 (0.01; 0.07)	165 (-391; 651)	4520	75%	24%	0%	1%	0.246	0.908	0.958	0.978
US	3L value set				0.04 (0.01; 0.07)	165 (-391; 651)	4613	74%	24%	1%	1%	0.246	0.895	0.945	0.972
	3L to 5L crosswalk				0.04 (0.01; 0.07)	165 (-391; 651)	4289	75%	24%	0%	1%	0.246	0.922	0.963	0.981
JP	3L value set				0.02 (-0.003; 0.05)	165 (-391; 651)	7193	72%	24%	0%	4%	0.246	0.773	0.860	0.910
	3L to 5L crosswalk				0.03 (0.01; 0.04)	165 (-391; 651)	6440	75%	24%	0%	1%	0.246	0.844	0.935	0.977
NL	3L value set	(5)	**Moderate depression**	Medium	0.09 (0.05; 0.12)	-152 (-724; 427)	-1753	29%	71%	0%	0%	0.709	0.999	1	1
	3L to 5L crosswalk				0.09 (0.07; 0.12)	-152 (-724; 427)	-1619	29%	71%	0%	0%	0.709	1	1	1
US	3L value set				0.06 (0.03; 0.08)	-152 (-724; 427)	-2565	29%	71%	0%	0%	0.709	0.999	0.999	0.999
	3L to 5L crosswalk				0.09 (0.06; 0.11)	-152 (-724; 427)	-1710	29%	71%	0%	0%	0.709	1	1	1
JP	3L value set				0.07 (0.04; 0.09)	-152 (-724; 427)	-2301	29%	71%	0%	0%	0.709	0.999	0.999	0.999
	3L to 5L crosswalk				0.07 (0.05; 0.08)	-152 (-724; 427)	-2260	29%	71%	0%	0%	0.709	1	1	1
NL	3L value set	(6)	**Moderate depression**	Large	0.13 (0.10; 0.16)	351 (-30; 738)	2705	96%	4%	0%	0%	0.036	1	1	1
	3L to 5L crosswalk				0.12 (0.10; 0.15)	351 (-30; 738)	2853	96%	4%	0%	0%	0.036	1	1	1
US	3L value set				0.11 (0.08; 0.13)	351 (-30; 738)	3240	96%	4%	0%	0%	0.036	1	1	1
	3L to 5L crosswalk				0.12 (0.1; 0.15)	351 (-30; 738)	2901	96%	4%	0%	0%	0.036	1	1	1
JP	3L value set				0.14 (0.11; 0.16)	351 (-30; 738)	2583	96%	4%	0%	0%	0.036	1	1	1
	3L to 5L crosswalk				0.10 (0.08; 0.11)	351 (-30; 738)	3526	96%	4%	0%	0%	0.036	1	1	1
NL	3L value set	(7)	**Severe depression**	Small	0.05 (0.01; 0.10)	96 (-310; 475)	1805	69%	31%	0%	0%	0.307	0.975	0.984	0.991
	3L to 5L crosswalk				0.04 (0.01; 0.08)	96 (-310; 475)	2208	69%	31%	0%	0%	0.307	0.975	0.987	0.993
US	3L value set				0.03 (0.002; 0.07)	96 (-310; 475)	2757	68%	31%	0%	1%	0.307	0.930	0.953	0.968
	3L to 5L crosswalk				0.04 (0.01; 0.07)	96 (-310; 475)	2405	69%	31%	0%	0%	0.307	0.965	0.980	0.987
JP	3L value set				0.03 (-0.0004; 0.05)	96 (-310; 475)	3497	67%	31%	0%	2%	0.307	0.896	0.932	0.955
	3L to 5L crosswalk				0.02 (0.01; 0.04)	96 (-310; 475)	4060	69%	31%	0%	0%	0.307	0.912	0.963	0.984

Country	Scoring method	Scenario	Patient population	Effect size	**Incremental QALYs** IQ (95% CI)	**Incremental costs** € (95% CI)	**ICER €/point**	**Distribution CE-plane (%)**				**Probability of cost-effectiveness**			
								NE	**SE**	**SW**	**NW**	**$p_{CE}(0)$**	**$p_{CE}(20000)$**	**$p_{CE}(30000)$**	**$p_{CE}(50000)$**
NL	3L value set	(8)	**Severe depression**	Medium	0.11 (0.06; 0.15)	-157 (-624; 292)	-1434	26%	74%	0%	0%	0.740	1	1	1
	3L to 5L crosswalk				0.08 (0.04; 0.11)	-157 (-624; 292)	-1980	26%	74%	0%	0%	0.740	1	1	1
US	3L value set				0.08 (0.05; 0.12)	-157 (-624; 292)	-1878	26%	74%	0%	0%	0.740	1	1	1
	3L to 5L crosswalk				0.08 (0.04; 0.11)	-157 (-624; 292)	-2076	26%	74%	0%	0%	0.740	1	1	1
JP	3L value set				0.08 (0.04; 0.11)	-157 (-624; 292)	-2059	26%	74%	0%	0%	0.740	1	1	1
	3L to 5L crosswalk				0.05 (0.03; 0.07)	-157 (-624; 292)	-3031	26%	74%	0%	0%	0.740	1	1	1
NL	3L value set	(9)	**Severe depression**	Large	0.24 (0.20; 0.29)	123 (-331; 592)	503	68%	32%	0%	0%	0.319	1	1	1
	3L to 5L crosswalk				0.18 (0.15; 0.22)	123 (-331; 592)	672	68%	32%	0%	0%	0.319	1	1	1
US	3L value set				0.17 (0.14; 0.21)	123 (-331; 592)	713	68%	32%	0%	0%	0.319	1	1	1
	3L to 5L crosswalk				0.17 (0.13; 0.21)	123 (-331; 592)	729	68%	32%	0%	0%	0.319	1	1	1
JP	3L value set				0.14 (0.11; 0.17)	123 (-331; 592)	895	68%	32%	0%	0%	0.319	1	1	1
	3L to 5L crosswalk				0.11 (0.09; 0.13)	123 (-331; 592)	1161	68%	32%	0%	0%	0.319	1	1	1
NL	3L value set	(10)	**Mild low back pain**	Small	0.03 (-0.01; 0.06)	131 (-386; 652)	4680	65%	30%	1%	4%	0.310	0.810	0.867	0.903
	3L to 5L crosswalk				0.03 (-0.0002; 0.05)	131 (-386; 652)	5057	67%	31%	0%	2%	0.310	0.827	0.893	0.936
US	3L value set				0.02 (-0.004; 0.05)	131 (-386; 652)	6159	64%	31%	1%	4%	0.310	0.770	0.833	0.886
	3L to 5L crosswalk				0.04 (0.02; 0.07)	131 (-386; 652)	3069	69%	31%	0%	0%	0.310	0.960	0.986	0.997
JP	3L value set				0.02 (0.002; 0.04)	131 (-386; 652)	6193	67%	31%	1%	1%	0.310	0.798	0.877	0.942
	3L to 5L crosswalk				0.03 (0.02; 0.05)	131 (-386; 652)	4158	69%	31%	0%	0%	0.310	0.934	0.984	0.999
NL	3L value set	(11)	**Mild low back pain**	Medium	0.10 (0.06; 0.13)	215 (-291; 713)	2263	80%	20%	0%	0%	0.197	1	1	1
	3L to 5L crosswalk				0.07 (0.04; 0.10)	215 (-291; 713)	3024	80%	20%	0%	0%	0.197	1	1	1
US	3L value set				0.07 (0.04; 0.10)	215 (-291; 713)	3083	80%	20%	0%	0%	0.197	0.998	0.999	1
	3L to 5L crosswalk				0.08 (0.05; 0.11)	215 (-291; 713)	2551	80%	20%	0%	0%	0.197	1	1	1
JP	3L value set				0.06 (0.03; 0.08)	215 (-291; 713)	3907	80%	20%	0%	0%	0.197	0.991	0.999	1
	3L to 5L crosswalk				0.05 (0.03; 0.07)	215 (-291; 713)	4158	80%	20%	0%	0%	0.197	0.994	1	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental **QALYs** IQ (95% CI)	Incremental **costs** € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	3L value set	(12)	**Mild low back pain**	Large	0.09 (0.07; 0.12)	380 (-90; 900)	4115	94%	6%	0%	0%	0.063	1	1	1
	3L to 5L crosswalk				0.07 (0.05; 0.09)	380 (-90; 900)	5476	94%	6%	0%	0%	0.063	1	1	1
US	3L value set				0.07 (0.06; 0.09)	380 (-90; 900)	5148	94%	6%	0%	0%	0.063	1	1	1
	3L to 5L crosswalk				0.09 (0.08; 0.12)	380 (-90; 900)	4010	94%	6%	0%	0%	0.063	1	1	1
JP	3L value set				0.06 (0.04; 0.07)	380 (-90; 900)	6663	94%	6%	0%	0%	0.063	0.996	1	1
	3L to 5L crosswalk				0.06 (0.05; 0.07)	380 (-90; 900)	6385	94%	6%	0%	0%	0.063	0.998	1	1
NL	3L value set	(13)	**Moderate low back pain**	Small	0.02 (-0.01; 0.05)	-97 (-668; 378)	-4147	35%	58%	4%	3%	0.620	0.919	0.936	0.941
	3L to 5L crosswalk				0.02 (0.001; 0.04)	-97 (-668; 378)	-4490	37%	60%	1%	2%	0.620	0.945	0.966	0.975
US	3L value set				0.02 (-0.004; 0.04)	-97 (-668; 378)	-5768	35%	58%	4%	3%	0.620	0.898	0.925	0.939
	3L to 5L crosswalk				0.03 (0.004; 0.05)	-97 (-668; 378)	-3418	37%	61%	1%	1%	0.620	0.973	0.981	0.988
JP	3L value set				0.02 (-0.003; 0.04)	-97 (-668; 378)	-5847	36%	58%	4%	2%	0.620	0.908	0.943	0.953
	3L to 5L crosswalk				0.02 (0.004; 0.04)	-97 (-668; 378)	-4926	38%	61%	1%	0%	0.620	0.952	0.980	0.993
NL	3L value set	(14)	**Moderate low back pain**	Medium	0.07 (0.04; 0.10)	291 (-165; 726)	4405	89%	11%	0%	0%	0.105	0.996	0.999	1
	3L to 5L crosswalk				0.06 (0.04; 0.08)	291 (-165; 726)	4904	89%	11%	0%	0%	0.105	0.998	1	1
US	3L value set				0.05 (0.03; 0.07)	291 (-165; 726)	5582	89%	11%	0%	0%	0.105	0.994	0.999	1
	3L to 5L crosswalk				0.08 (0.05; 0.10)	291 (-165; 726)	3873	89%	11%	0%	0%	0.105	1	1	1
JP	3L value set				0.05 (0.03; 0.07)	291 (-165; 726)	5769	89%	11%	0%	0%	0.105	0.994	0.999	1
	3L to 5L crosswalk				0.05 (0.03; 0.06)	291 (-165; 726)	6009	89%	11%	0%	0%	0.105	0.996	1	1
NL	3L value set	(15)	**Moderate low back pain**	Large	0.11 (0.08; 0.14)	98 (-327; 559)	936	67%	33%	0%	0%	0.324	1	1	1
	3L to 5L crosswalk				0.1 (0.09; 0.12)	98 (-327; 559)	943	67%	33%	0%	0%	0.324	1	1	1
US	3L value set				0.09 (0.07; 0.11)	98 (-327; 559)	1111	67%	33%	0%	0%	0.324	1	1	1
	3L to 5L crosswalk				0.12 (0.10; 0.14)	98 (-327; 559)	809	67%	33%	0%	0%	0.324	1	1	1
JP	3L value set				0.11 (0.09; 0.13)	98 (-327; 559)	921	67%	33%	0%	0%	0.324	1	1	1
	3L to 5L crosswalk				0.08 (0.07; 0.10)	98 (-327; 559)	1162	67%	33%	0%	0%	0.324	1	1	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental **QALYs** IQ (95% CI)	Incremental **costs** € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	**SE**	**SW**	**NW**	**$p_{CE}(0)$**	**$p_{CE}(20000)$**	**$p_{CE}(30000)$**	**$p_{CE}(50000)$**
NL	3L value set	(16)	**Severe low back pain**	Small	0.04 (-0.004; 0.09)	132 (-398; 646)	3096	66%	31%	0%	3%	0.315	0.898	0.926	0.944
	3L to 5L crosswalk				0.03 (-0.01; 0.06)	132 (-398; 646)	5272	63%	30%	1%	6%	0.315	0.791	0.850	0.890
US	3L value set				0.03 (-0.004; 0.07)	132 (-398; 646)	4183	66%	31%	0%	3%	0.315	0.857	0.908	0.939
	3L to 5L crosswalk				0.03 (-0.003; 0.06)	132 (-398; 646)	4390	65%	31%	0%	3%	0.315	0.850	0.901	0.934
JP	3L value set				0.04 (0.01; 0.07)	132 (-398; 646)	3561	68%	31%	0%	1%	0.315	0.928	0.969	0.984
	3L to 5L crosswalk				0.02 (0.002; 0.04)	132 (-398; 646)	6839	67%	31%	0%	2%	0.315	0.776	0.870	0.939
NL	3L value set	(17)	**Severe low back pain**	Medium	0.13 (0.09; 0.18)	-39 (-481; 371)	-297	44%	56%	0%	0%	0.561	1	1	1
	3L to 5L crosswalk				0.10 (0.07; 0.13)	-39 (-481; 371)	-386	44%	56%	0%	0%	0.561	1	1	1
US	3L value set				0.08 (0.05; 0.12)	-39 (-481; 371)	-464	44%	56%	0%	0%	0.561	1	1	1
	3L to 5L crosswalk				0.09 (0.06; 0.12)	-39 (-481; 371)	-451	44%	56%	0%	0%	0.561	1	1	1
JP	3L value set				0.07 (0.04; 0.10)	-39 (-481; 371)	-562	44%	56%	0%	0%	0.561	1	1	1
	3L to 5L crosswalk				0.05 (0.03; 0.06)	-39 (-481; 371)	-863	44%	56%	0%	0%	0.561	1	1	1
NL	3L value set	(18)	**Severe low back pain**	Large	0.14 (0.11; 0.18)	113 (-371; 604)	780	67%	33%	0%	0%	0.324	1	1	1
	3L to 5L crosswalk				0.12 (0.10; 0.15)	113 (-371; 604)	936	67%	33%	0%	0%	0.324	1	1	1
US	3L value set				0.12 (0.09; 0.14)	113 (-371; 604)	976	67%	33%	0%	0%	0.324	1	1	1
	3L to 5L crosswalk				0.13 (0.11; 0.16)	113 (-371; 604)	837	67%	33%	0%	0%	0.324	1	1	1
JP	3L value set				0.13 (0.10; 0.15)	113 (-371; 604)	903	67%	33%	0%	0%	0.324	1	1	1
	3L to 5L crosswalk				0.08 (0.07; 0.10)	113 (-371; 604)	1352	67%	33%	0%	0%	0.324	1	1	1
NL	3L value set	(19)	**Mild osteoarthritis**	Small	0.03 (-0.01; 0.07)	127 (-314; 613)	3973	67%	29%	1%	3%	0.301	0.864	0.908	0.940
	3L to 5L crosswalk				0.03 (0.002; 0.06)	127 (-314; 613)	4506	68%	30%	0%	2%	0.301	0.874	0.928	0.965
US	3L value set				0.02 (-0.01; 0.05)	s127 (-314; 613)	5589	65%	29%	1%	5%	0.301	**0.798**	**0.853**	0.893
	3L to 5L crosswalk				0.01 (-0.01; 0.04)	127 (-314; 613)	8765	58%	27%	3%	2%	0.301	**0.662**	**0.726**	0.785
JP	3L value set				0.02 (-0.001; 0.04)	127 (-314; 613)	5956	68%	29%	1%	2%	0.301	**0.800**	**0.880**	0.932
	3L to 5L crosswalk				0.003 (-0.01; 0.02)	127 (-314; 613)	44792	43%	21%	9%	27%	0.301	**0.426**	**0.726**	0.785

Country	Scoring method	Scenario	Patient population	Effect size	**Incremental QALYs** IQ (95% CI)	**Incremental costs** € (95% CI)	**ICER €/point**	**Distribution CE-plane (%)**				**Probability of cost-effectiveness**			
								NE	**SE**	**SW**	**NW**	**p_{CE}(0)**	**p_{CE}(20000)**	**p_{CE}(30000)**	**p_{CE}(50000)**
NL	3L value set	(20)	**Mild osteoarthritis**	Medium	0.04 (0.03; 0.06)	-258 (-788; 193)	-5973	16%	84%	0%	0%	0.845	0.999	1	1
	3L to 5L crosswalk				0.04 (0.02; 0.05)	-258 (-788; 193)	-6848	16%	84%	0%	0%	0.845	0.999	1	1
US	3L value set				0.04 (0.02; 0.05)	-258 (-788; 193)	-6948	16%	84%	0%	0%	0.845	0.999	1	1
	3L to 5L crosswalk				0.03 (0.01; 0.05)	-258 (-788; 193)	-9061	16%	84%	0%	0%	0.845	0.995	0.997	0.998
JP	3L value set				0.05 (0.03; 0.07)	-258 (-788; 193)	-5126	16%	84%	0%	0%	0.845	1	1	1
	3L to 5L crosswalk				0.02 (0.01; 0.03)	-258 (-788; 193)	-13888	16%	84%	0%	0%	0.845	0.984	0.992	0.996
NL	3L value set	(21)	**Mild osteoarthritis**	Large	0.06 (0.04; 0.07)	366 (-61; 806)	6286	95%	5%	0%	0%	0.045	0.998	1	1
	3L to 5L crosswalk				0.06 (0.04; 0.07)	366 (-61; 806)	6497	95%	5%	0%	0%	0.045	0.999	1	1
US	3L value set				0.06 (0.04; 0.07)	366 (-61; 806)	6593	95%	5%	0%	0%	0.045	0.997	1	1
	3L to 5L crosswalk				0.04 (0.03; 0.06)	366 (-61; 806)	8754	95%	5%	0%	0%	0.045	0.952	0.998	1
JP	3L value set				0.09 (0.08; 0.11)	366 (-61; 806)	4001	95%	5%	0%	0%	0.045	1	1	1
	3L to 5L crosswalk				0.04 (0.02; 0.05)	366 (-61; 806)	9955	95%	5%	0%	0%	0.045	0.919	0.966	0.999
NL	3L value set	(22)	**Moderate osteoarthritis**	Small	0.02 (-0.01; 0.05)	-97 (-668; 378)	-4147	35%	58%	4%	3%	0.620	0.919	0.936	0.941
	3L to 5L crosswalk				0.02 (0.001; 0.04)	-97 (-668; 378)	-4490	37%	60%	2%	1%	0.620	0.945	0.966	0.975
US	3L value set				0.02 (-0.004; 0.04)	-97 (-668; 378)	-5768	35%	58%	4%	3%	0.620	0.898	0.925	0.939
	3L to 5L crosswalk				0.03 (0.004; 0.05)	-97 (-668; 378)	-3418	37%	61%	1%	1%	0.620	0.973	0.981	0.988
JP	3L value set				0.02 (-0.003; 0.04)	-97 (-668; 378)	-5847	36%	58%	4%	2%	0.620	0.908	0.943	0.953
	3L to 5L crosswalk				0.02 (0.004; 0.04)	-97 (-668; 378)	-4926	38%	61%	1%	0%	0.620	0.952	0.980	0.993
NL	3L value set	(23)	**Moderate osteoarthritis**	Medium	0.07 (0.04; 0.10)	291 (-165; 726)	4405	89%	11%	0%	0%	0.105	0.996	0.999	1
	3L to 5L crosswalk				0.06 (0.04; 0.08)	291 (-165; 726)	4904	89%	11%	0%	0%	0.105	0.998	1	1
US	3L value set				0.05 (0.03; 0.07)	291 (-165; 726)	5581	89%	11%	0%	0%	0.105	0.994	0.999	1
	3L to 5L crosswalk				0.08 (0.05; 0.10)	291 (-165; 726)	3873	89%	11%	0%	0%	0.105	1	1	1
JP	3L value set				0.05 (0.03; 0.07)	291 (-165; 726)	5769	89%	11%	0%	0%	0.105	0.994	0.999	1
	3L to 5L crosswalk				0.05 (0.03; 0.06)	291 (-165; 726)	6009	89%	11%	0%	0%	0.105	0.996	1	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental **QALYs** IQ (95% CI)	Incremental **costs** € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	3L value set	(24)	**Moderate osteoarthritis**	Large	0.14 (0.11; 0.17)	422 (-22; 890)	3048	96%	4%	0%	0%	0.035	1	1	1
	3L to 5L crosswalk				0.11 (0.09; 0.14)	422 (-22; 890)	3741	96%	4%	0%	0%	0.035	1	1	1
US	3L value set				0.12 (0.09; 0.14)	422 (-22; 890)	3597	96%	4%	0%	0%	0.035	1	1	1
	3L to 5L crosswalk				0.13 (0.11; 0.16)	422 (-22; 890)	3141	96%	4%	0%	0%	0.035	1	1	1
JP	3L value set				0.10 (0.08; 0.12)	422 (-22; 890)	4337	96%	4%	0%	0%	0.035	1	1	1
	3L to 5L crosswalk				0.09 (0.07; 0.10)	422 (-22; 890)	4945	96%	4%	0%	0%	0.035	1	1	1
NL	3L value set	(25)	**Severe osteoarthritis**	Small	0.04 (-0.004; 0.09)	132 (-399; 646)	3097	66%	31%	0%	3%	0.315	0.898	0.926	0.945
	3L to 5L crosswalk				0.03 (-0.01; 0.06)	132 (-399; 646)	5272	63%	30%	1%	6%	0.315	0.791	0.850	0.890
US	3L value set				0.03 (-0.004; 0.07)	132 (-399; 646)	4183	66%	31%	0%	3%	0.315	0.858	0.908	0.939
	3L to 5L crosswalk				0.03 (-0.003; 0.06)	132 (-399; 646)	4389	65%	32%	0%	3%	0.315	0.850	0.901	0.934
JP	3L value set				0.04 (0.01; 0.07)	132 (-399; 646)	3561	68%	31%	0%	1%	0.315	0.928	0.969	0.984
	3L to 5L crosswalk				0.02 (0.002; 0.04)	132 (-399; 646)	6839	67%	31%	0%	2%	0.315	0.776	0.870	0.939
NL	3L value set	(26)	**Severe osteoarthritis**	Medium	0.07 (0.04; 0.09)	-430 (-890; -37)	-6329	2%	98%	0%	0%	0.982	1	1	1
	3L to 5L crosswalk				0.02 (-0.002; 0.03)	-430 (-890; -37)	-26515	1%	94%	5%	0%	0.982	0.994	0.993	0.988
US	3L value set				0.05 (0.03; 0.07)	-430 (-890; -37)	-8861	2%	98%	0%	0%	0.982	1	1	1
	3L to 5L crosswalk				0.02 (0.01; 0.04)	-430 (-890; -37)	-17214	2%	98%	0%	0%	0.982	0.999	0.999	0.999
JP	3L value set				0.05 (0.03; 0.07)	-430 (-890; -37)	-9060	2%	98%	0%	0%	0.982	1	1	1
	3L to 5L crosswalk				0.01 (0.005; 0.02)	-430 (-890; -37)	-31030	2%	98%	0%	0%	0.982	0.998	0.998	0.999
NL	3L value set	(27)	**Severe osteoarthritis**	Large	0.13 (0.10; 0.17)	357 (-95; 796)	2650	95%	5%	0%	0%	0.050	1	1	1
	3L to 5L crosswalk				0.09 (0.06; 0.11)	357 (-95; 796)	4084	95%	5%	0%	0%	0.050	1	1	1
US	3L value set				0.10 (0.08; 0.13)	357 (-95; 796)	3526	95%	5%	0%	0%	0.050	1	1	1
	3L to 5L crosswalk				0.10 (0.08; 0.12)	357 (-95; 796)	3595	95%	5%	0%	0%	0.050	1	1	1
JP	3L value set				0.09 (0.06; 0.11)	357 (-95; 796)	4141	95%	5%	0%	0%	0.050	1	1	1
	3L to 5L crosswalk				0.05 (0.04; 0.06)	357 (-95; 796)	7222	95%	5%	0%	0%	0.050	0.989	1	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental **QALYs** IQ (95% CI)	Incremental **costs** € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	3L value set	(28)	**Mild cancer**	Small	0.03 (-0.003; 0.05)	178 (-293; 677)	6517	74%	23%	10%	3%	0.236	**0.826**	0.889	0.930
	3L to 5L crosswalk				0.02 (-0.01; 0.04)	178 (-293; 677)	11279	71%	22%	2%	5%	0.236	**0.659**	0.759	0.845
US	3L value set				0.03 (0.01; 0.05)	178 (-293; 677)	5804	76%	23%	0%	1%	0.236	0.882	0.944	0.969
	3L to 5L crosswalk				0.02 (0.003; 0.05)	178 (-293; 677)	7165	75%	23%	0%	1%	0.236	0.819	0.903	0.952
JP	3L value set				0.02 (-0.0004; 0.04)	178 (-293; 677)	9007	74%	23%	1%	2%	0.236	0.743	**0.842**	0.916
	3L to 5L crosswalk				0.01 (-0.003; 0.03)	178 (-293; 677)	14832	71%	22%	2%	5%	0.236	0.573	**0.699**	0.816
NL	3L value set	(29)	**Mild cancer**	Medium	0.03 (0.02; 0.04)	41 (-401; 492)	1408	56%	44%	0%	0%	0.436	0.979	0.995	0.999
	3L to 5L crosswalk				0.03 (0.01; 0.04)	41 (-401; 492)	1540	56%	44%	0%	0%	0.436	0.973	0.992	0.999
US	3L value set				0.03 (0.02; 0.05)	41 (-401; 492)	1236	56%	44%	0%	0%	0.436	0.990	0.998	1
	3L to 5L crosswalk				0.02 (0.01; 0.04)	41 (-401; 492)	1681	56%	44%	0%	0%	0.436	0.963	0.990	0.999
JP	3L value set				0.05 (0.03; 0.07)	41 (-401; 492)	821	56%	44%	0%	0%	0.436	0.999	1	1
	3L to 5L crosswalk				0.02 (0.01; 0.03)	41 (-401; 492)	1906	56%	44%	0%	0%	0.436	0.944	0.985	0.997
NL	3L value set	(30)	**Mild cancer**	Large	0.05 (0.04; 0.06)	53 (-396; 502)	1100	58%	42%	0%	0%	0.421	1	1	1
	3L to 5L crosswalk				0.05 (0.04; 0.06)	53 (-396; 502)	1124	58%	42%	0%	0%	0.421	1	1	1
US	3L value set				0.06 (0.04; 0.07)	53 (-396; 502)	964	58%	42%	0%	0%	0.421	1	1	1
	3L to 5L crosswalk				0.05 (0.04; 0.06)	53 (-396; 502)	1026	58%	42%	0%	0%	0.421	1	1	1
JP	3L value set				0.09 (0.07; 0.10)	53 (-396; 502)	618	58%	42%	0%	0%	0.421	1	1	1
	3L to 5L crosswalk				0.05 (0.04; 0.06)	53 (-396; 502)	1156	58%	42%	0%	0%	0.421	1	1	1
NL	3L value set	(31)	**Moderate cancer**	Small	0.03 (-0.004; 0.07)	285 (-121; 712)	8717	86%	10%	0%	4%	0.101	0.801	0.872	0.909
	3L to 5L crosswalk				0.05 (0.02; 0.08)	285 (-121; 712)	6083	90%	10%	0%	0%	0.101	0.952	0.984	0.995
US	3L value set				0.03 (-0.003; 0.06)	285 (-121; 712)	10775	86%	10%	0%	4%	0.101	**0.745**	0.845	0.908
	3L to 5L crosswalk				0.06 (0.03; 0.09)	285 (-121; 712)	4768	90%	10%	0%	0%	0.101	**0.989**	0.997	1
JP	3L value set				0.03 (0.002; 0.05)	285 (-121; 712)	11054	89%	10%	0%	1%	0.101	**0.756**	0.868	0.934
	3L to 5L crosswalk				0.04 (0.02; 0.06)	285 (-121; 712)	6609	90%	10%	0%	0%	0.101	**0.978**	0.998	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental **QALYs** IQ (95% CI)	Incremental **costs** € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	3L value set	(32)	**Moderate cancer**	Medium	0.07 (0.04; 0.10)	-174 (-628; 281)	-2534	22%	78%	0%	0%	0.776	1	1	1
	3L to 5L crosswalk				0.08 (0.06; 0.11)	-174 (-628; 281)	-2053	22%	78%	0%	0%	0.776	1	1	1
US	3L value set				0.06 (0.03; 0.08)	-174 (-628; 281)	-3111	22%	78%	0%	0%	0.776	0.999	1	1
	3L to 5L crosswalk				0.09 (0.06; 0.11)	-174 (-628; 281)	-1978	22%	78%	0%	0%	0.776	1	1	1
JP	3L value set				0.08 (0.06; 0.11)	-174 (-628; 281)	-2062	22%	78%	0%	0%	0.776	1	1	1
	3L to 5L crosswalk				0.07 (0.05; 0.09)	-174 (-628; 281)	-2539	22%	78%	0%	0%	0.776	1	1	1
NL	3L value set	(33)	**Moderate cancer**	Large	0.10 (0.08; 0.13)	94 (-329; 514)	940	68%	32%	0%	0%	0.314	1	1	1
	3L to 5L crosswalk				0.11 (0.09; 0.13)	94 (-329; 514)	885	68%	32%	0%	0%	0.314	1	1	1
US	3L value set				0.09 (0.07; 0.11)	94 (-329; 514)	1035	68%	32%	0%	0%	0.314	1	1	1
	3L to 5L crosswalk				0.11 (0.09; 0.14)	94 (-329; 514)	830	68%	32%	0%	0%	0.314	1	1	1
JP	3L value set				0.12 (0.10; 0.14)	94 (-329; 514)	797	68%	32%	0%	0%	0.314	1	1	1
	3L to 5L crosswalk				0.09 (0.08; 0.11)	94 (-329; 514)	1036	68%	32%	0%	0%	0.314	1	1	1
NL	3L value set	(34)	**Severe cancer**	Small	0.05 (-0.002; 0.11)	364 (-119; 965)	6737	89%	9%	0%	2%	0.085	0.872	0.915	0.945
	3L to 5L crosswalk				0.05 (0.01; 0.10)	364 (-119; 965)	6868	90%	9%	0%	1%	0.085	0.891	0.942	0.967
US	3L value set				0.04 (-0.003; 0.09)	364 (-119; 965)	8414	88%	9%	0%	3%	0.085	0.826	0.881	0.924
	3L to 5L crosswalk				0.06 (0.01; 0.11)	364 (-119; 965)	5948	91%	8%	0%	1%	0.085	0.920	0.957	0.975
JP	3L value set				0.04 (-0.003; 0.08)	364 (-119; 965)	9082	88%	8%	0%	4%	0.085	0.800	0.870	0.916
	3L to 5L crosswalk				0.04 (0.01; 0.07)	364 (-119; 965)	9782	91%	8%	0%	1%	0.085	0.818	0.903	0.959
NL	3L value set	(35)	**Severe cancer**	Medium	0.14 (0.09; 0.20)	109 (-409; 617)	775	66%	34%	0%	0%	0.336	1	1	1
	3L to 5L crosswalk				0.11 (0.07; 0.16)	109 (-409; 617)	958	66%	34%	0%	0%	0.336	1	1	1
US	3L value set				0.11 (0.06; 0.16)	109 (-409; 617)	998	66%	34%	0%	0%	0.336	1	1	1
	3L to 5L crosswalk				0.13 (0.07; 0.18)	109 (-409; 617)	871	66%	34%	0%	0%	0.336	1	1	1
JP	3L value set				0.10 (0.05; 0.14)	109 (-409; 617)	1143	66%	34%	0%	0%	0.336	1	1	1
	3L to 5L crosswalk				0.08 (0.05; 0.11)	109 (-409; 617)	1391	66%	34%	0%	0%	0.336	1	1	1

Country	Scoring method	Scenario	Patient population	Effect size	**Incremental QALYs** IQ (95% CI)	**Incremental costs** € (95% CI)	**ICER €/point**	**Distribution CE-plane (%)**				**Probability of cost-effectiveness**			
								NE	**SE**	**SW**	**NW**	**p_{CE}(0)**	**p_{CE}(20000)**	**p_{CE}(30000)**	**p_{CE}(50000)**
NL	3L value set	(36)	**Severe cancer**	Large	0.24 (0.18; 0.29)	-217 (-690; 238)	-921	18%	82%	0%	0%	0.825	1	1	1
	3L to 5L crosswalk				0.18 (0.14; 0.23)	-217 (-690; 238)	-1180	18%	82%	0%	0%	0.825	1	1	1
US	3L value set				0.19 (0.15; 0.24)	-217 (-690; 238)	-1134	18%	82%	0%	0%	0.825	1	1	1
	3L to 5L crosswalk				0.20 (0.15; 0.25)	-217 (-690; 238)	-1108	18%	82%	0%	0%	0.825	1	1	1
JP	3L value set				0.16 (0.12; 0.21)	-217 (-690; 238)	-1323	18%	82%	0%	0%	0.825	1	1	1
	3L to 5L crosswalk				0.12 (0.09; 0.15)	-217 (-690; 238)	-1775	18%	82%	0%	0%	0.825	1	1	1

NL: the Netherlands. US: United States. JP: Japan. QALY: quality-adjusted life-year. IQ: incremental QALY. 95% CI: 95% confidence interval. ICER: Incremental Cost-Effectiveness Ratio. NE: northeast. SE: southeast. SW: southwest. NW: northwest. p_{CE}: probability of the intervention being cost-effective compared to control.

Supplementary Table 4.2 | Cost-utility analysis results for 5L value set and 5L to 3L crosswalk per country

Country	Scoring method	Scenario	Patient population	Effect size	Incremental QALYs IQ (95% CI)	Incremental costs € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	p_{CE}(0)	p_{CE}(20000)	p_{CE}(30000)	p_{CE}(50000)
NL	5L value set	(1)	**Mild depression**	Small	0.02 (-0.02; 0.06)	195 (-320; 742)	8552	64%	22%	2%	12%	0.242	0.696	0.758	0.801
	5L to 3L crosswalk				0.02 (-0.01; 0.06)	195 (-320; 742)	8937	65%	22%	2%	11%	0.242	0.699	0.757	0.808
US	5L value set				0.03 (-0.01; 0.08)	195 (-320; 742)	6221	69%	23%	1%	7%	0.242	**0.790**	0.833	0.868
	5L to 3L crosswalk				0.02 (-0.01; 0.05)	195 (-320; 742)	8447	70%	23%	1%	6%	0.242	**0.743**	0.817	0.867
JP	5L value set				0.01 (-0.03; 0.05)	195 (-320; 742)	15705	54%	19%	5%	22%	0.242	**0.537**	**0.600**	**0.658**
	5L to 3L crosswalk				0.02 (-0.001; 0.05)	195 (-320; 742)	8107	73%	23%	1%	3%	0.242	**0.782**	**0.861**	**0.821**
NL	5L value set	(2)	**Mild depression**	Medium	0.11 (0.08; 0.15)	-228 (-692; 295)	-2049	19%	81%	0%	0%	0.809	1	1	1
	5L to 3L crosswalk				0.12 (0.09; 0.15)	-228 (-692; 295)	-1902	19%	81%	0%	0%	0.809	1	1	1
US	5L value set				0.13 (0.09; 0.17)	-228 (-692; 295)	-1800	19%	81%	0%	0%	0.809	1	1	1
	5L to 3L crosswalk				0.11 (0.08; 0.13)	-228 (-692; 295)	-2151	19%	81%	0%	0%	0.809	1	1	1
JP	5L value set				0.08 (0.05; 0.12)	-228 (-692; 295)	-2858	19%	81%	0%	0%	0.809	1	1	1
	5L to 3L crosswalk				0.09 (0.07; 0.11)	-228 (-692; 295)	-2470	19%	81%	0%	0%	0.809	1	1	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental QALYs IQ (95% CI)	Incremental costs € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	5L value set	(3)	**Mild depression**	Large	0.18 (0.14; 0.21)	274 (-217; 724)	1557	89%	11%	0%	0%	0.113	1	1	1
	5L to 3L crosswalk				0.19 (0.16; 0.22)	274 (-217; 724)	1440	89%	11%	0%	0%	0.113	1	1	1
US	5L value set				0.18 (0.14; 0.22)	274 (-217; 724)	1511	89%	11%	0%	0%	0.113	1	1	1
	5L to 3L crosswalk				0.17 (0.14; 0.19)	274 (-217; 724)	1627	89%	11%	0%	0%	0.113	1	1	1
JP	5L value set				0.14 (0.1; 0.17)	274 (-217; 724)	2018	89%	11%	0%	0%	0.113	1	1	1
	5L to 3L crosswalk				0.16 (0.14; 0.18)	274 (-217; 724)	1687	89%	11%	0%	0%	0.113	1	1	1
NL	5L value set	(4)	**Moderate depression**	Small	0.05 (0.003; 0.08)	43 (-448; 524)	952	55%	43%	1%	1%	0.439	0.952	0.968	0.975
	5L to 3L crosswalk				0.07 (0.03; 0.10)	43 (-448; 524)	628	56%	44%	0%	0%	0.439	0.995	0.999	0.999
US	5L value set				0.05 (0.003; 0.09)	43 (-448; 524)	898	55%	43%	1%	1%	0.439	0.960	0.972	0.979
	5L to 3L crosswalk				0.06 (0.03; 0.09)	43 (-448; 524)	713	56%	44%	0%	0%	0.439	0.997	1	1
JP	5L value set				0.02 (-0.02; 0.06)	43 (-448; 524)	2236	47%	38%	6%	9%	0.439	0.772	0.806	0.827
	5L to 3L crosswalk				0.04 (0.02; 0.06)	43 (-448; 524)	1093	56%	44%	0%	0%	0.439	0.986	0.994	0.999
NL	5L value set	(5)	**Moderate depression**	Medium	0.13 (0.09; 0.16)	-296 (-813; 193)	-2350	13%	87%	0%	0%	0.873	1	1	1
	5L to 3L crosswalk				0.14 (0.10; 0.17)	-296 (-813; 193)	-2141	13%	87%	0%	0%	0.873	1	1	1
US	5L value set				0.13 (0.09; 0.17)	-296 (-813; 193)	-2248	13%	87%	0%	0%	0.873	1	1	1
	5L to 3L crosswalk				0.12 (0.09; 0.14)	-296 (-813; 193)	-2516	13%	87%	0%	0%	0.873	1	1	1
JP	5L value set				0.08 (0.05; 0.12)	-296 (-813; 193)	-3526	13%	87%	0%	0%	0.873	1	1	1
	5L to 3L crosswalk				0.10 (0.08; 0.12)	-296 (-813; 193)	-2896	13%	87%	0%	0%	0.873	1	1	1
NL	5L value set	(6)	**Moderate depression**	Large	0.19 (0.15; 0.22)	-110 (-493; 312)	-590	29%	71%	0%	0%	0.709	1	1	1
	5L to 3L crosswalk				0.20 (0.17; 0.23)	-110 (-493; 312)	-551	29%	71%	0%	0%	0.709	1	1	1
US	5L value set				0.20 (0.16; 0.24)	-110 (-493; 312)	-559	29%	71%	0%	0%	0.709	1	1	1
	5L to 3L crosswalk				0.18 (0.15; 0.2)	-110 (-493; 312)	-622	29%	71%	0%	0%	0.709	1	1	1
JP	5L value set				0.14 (0.1; 0.18)	-110 (-493; 312)	-770	29%	71%	0%	0%	0.709	1	1	1
	5L to 3L crosswalk				0.17 (0.15; 0.18)	-110 (-493; 312)	-664	29%	71%	0%	0%	0.709	1	1	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental QALYs IQ (95% CI)	Incremental costs € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	5L value set	(7)	**Severe depression**	Small	0.04 (-0.01; 0.09)	68 (-448; 571)	1810	56%	37%	2%	5%	0.391	0.867	0.893	0.912
	5L to 3L crosswalk				0.03 (-0.01; 0.07)	68 (-448; 571)	2391	55%	36%	3%	6%	0.391	0.830	0.862	0.886
US	5L value set				0.05 (-0.003; 0.10)	68 (-448; 571)	1373	58%	38%	1%	3%	0.391	0.927	0.941	0.952
	5L to 3L crosswalk				0.03 (-0.01; 0.06)	68 (-448; 571)	2618	57%	37%	2%	4%	0.391	0.848	0.886	0.914
JP	5L value set				0.04 (-0.004; 0.09)	68 (-448; 571)	1523	58%	38%	1%	3%	0.391	0.922	0.937	0.949
	5L to 3L crosswalk				0.02 (-0.003; 0.05)	68 (-448; 571)	2883	58%	38%	1%	3%	0.391	0.848	0.901	0.932
NL	5L value set	(8)	**Severe depression**	Medium	0.12 (0.07; 0.17)	-378 (-919; 140)	-3130	7%	93%	0%	0%	.933	1	1	1
	5L to 3L crosswalk				0.10 (0.06; 0.14)	-378 (-919; 140)	-3619	7%	93%	0%	0%	.933	1	1	1
US	5L value set				0.14 (0.09; 0.19)	-378 (-919; 140)	-2726	7%	93%	0%	0%	.933	1	1	1
	5L to 3L crosswalk				0.09 (0.06; 0.12)	-378 (-919; 140)	-4111	7%	93%	0%	0%	.933	1	1	1
JP	5L value set				0.11 (0.06; 0.15)	-378 (-919; 140)	-3546	7%	93%	0%	0%	.933	1	1	1
	5L to 3L crosswalk				0.08 (0.05; 0.1)	-378 (-919; 140)	-4739	7%	93%	0%	0%	.933	1	1	1
NL	5L value set	(9)	**Severe depression**	Large	0.21 (0.16; 0.25)	83 (-405; 581)	406	65%	35%	0%	0%	0.354	1	1	1
	5L to 3L crosswalk				0.20 (0.17; 0.24)	83 (-405; 581)	413	65%	35%	0%	0%	0.354	1	1	1
US	5L value set				0.21 (0.16; 0.26)	83 (-405; 581)	391	65%	35%	0%	0%	0.354	1	1	1
	5L to 3L crosswalk				0.18 (0.15; 0.2)	83 (-405; 581)	477	65%	35%	0%	0%	0.354	1	1	1
JP	5L value set				0.18 (0.13; 0.22)	83 (-405; 581)	472	65%	35%	0%	0%	0.354	1	1	1
	5L to 3L crosswalk				0.17 (0.14; 0.19)	83 (-405; 581)	500	65%	35%	0%	0%	0.354	1	1	1
NL	5L value set	(10)	**Mild low back pain**	Small	0.03 (-0.01; 0.78)	460 (-6; 1041)	13849	90%	3%	0%	7%	0.034	0.641	0.764	0.845
	5L to 3L crosswalk				0.03 (-0.002; 0.07)	460 (-6; 1041)	13646	93%	3%	1%	3%	0.034	0.678	0.814	0.895
US	5L value set				0.05 (0.01; 0.10)	460 (-6; 1041)	9019	95%	3%	0%	2%	0.034	0.835	0.912	0.954
	5L to 3L crosswalk				0.04 (0.01; 0.07)	460 (-6; 1041)	12013	96%	3%	0%	1%	0.034	0.773	0.902	0.973
JP	5L value set				0.03 (-0.01; 0.07)	460 (-6; 1041)	17604	85%	3%	1%	11%	0.034	**0.549**	**0.668**	0.775
	5L to 3L crosswalk				0.04 (0.02; 0.06)	460 (-6; 1041)	11345	96%	4%	0%	0%	0.034	**0.841**	**0.958**	0.996

Country	Scoring method	Scenario	Patient population	Effect size	Incremental QALYs IQ (95% CI)	Incremental costs € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	5L value set	(11)	**Mild low back pain**	Medium	0.11 (0.06; 0.15)	393 (-134; 953)	3703	92%	8%	0%	0%	0.075	0.998	0.999	1
	5L to 3L crosswalk				0.12 (0.09; 0.15)	393 (-134; 953)	3277	92%	8%	0%	0%	0.075	1	1	1
US	5L value set				0.13 (0.09; 0.18)	393 (-134; 953)	2944	92%	8%	0%	0%	0.075	1	1	1
	5L to 3L crosswalk				0.12 (0.09; 0.14)	393 (-134; 953)	3377	92%	8%	0%	0%	0.075	1	1	1
JP	5L value set				0.10 (0.06; 0.14)	393 (-134; 953)	4138	92%	8%	0%	0%	0.075	0.997	1	1
	5L to 3L crosswalk				0.12 (0.10; 0.14)	393 (-134; 953)	3364	92%	8%	0%	0%	0.075	0.997	1	1
NL	5L value set	(12)	**Mild low back pain**	Large	0.21 (0.17; 0.26)	225 (-221; 756)	1057	82%	18%	0%	0%	0.174	1	1	1
	5L to 3L crosswalk				0.25 (0.21; 0.28)	225 (-221; 756)	914	82%	18%	0%	0%	0.174	1	1	1
US	5L value set				0.24 (0.20; 0.29)	225 (-221; 756)	933	82%	18%	0%	0%	0.174	1	1	1
	5L to 3L crosswalk				0.24 (0.21; 0.26)	225 (-221; 756)	951	82%	18%	0%	0%	0.174	1	1	1
JP	5L value set				0.18 (0.14; 0.22)	225 (-221; 756)	1265	82%	18%	0%	0%	0.174	1	1	1
	5L to 3L crosswalk				0.22 (0.20; 0.24)	225 (-221; 756)	1039	82%	18%	0%	0%	0.174	1	1	1
NL	5L value set	(13)	**Moderate low back pain**	Small	0.02 (-0.02; 0.06)	233 (-186; 703)	11662	70%	13%	2%	15%	0.152	**0.642**	**0.711**	0.764
	5L to 3L crosswalk				0.04 (0.01; 0.08)	233 (-186; 703)	5250	84%	15%	0%	1%	0.152	**0.939**	**0.962**	0.984
US	5L value set				0.04 (-0.01; 0.08)	233 (-186; 703)	5845	81%	15%	1%	3%	0.152	0.864	0.910	0.935
	5L to 3L crosswalk				0.04 (0.01; 0.06)	233 (-186; 703)	5707	85%	15%	0%	0%	0.152	0.949	0.979	0.994
JP	5L value set				0.02 (-0.02; 0.06)	233 (-186; 703)	11900	70%	13%	2%	15%	0.152	**0.630**	**0.705**	0.762
	5L to 3L crosswalk				0.04 (0.02; 0.05)	233 (-186; 703)	6540	85%	15%	0%	0%	0.152	**0.939**	**0.982**	0.998
NL	5L value set	(14)	**Moderate low back pain**	Medium	0.12 (0.07; 0.16)	39 (-416; 488)	330	56%	44%	0%	0%	0.435	1	1	1
	5L to 3L crosswalk				0.14 (0.10; 0.17)	39 (-416; 488)	281	56%	44%	0%	0%	0.435	1	1	1
US	5L value set				0.15 (0.10; 0.20)	39 (-416; 488)	258	56%	44%	0%	0%	0.435	1	1	1
	5L to 3L crosswalk				0.13 (0.10; 0.15)	39 (-416; 488)	308	56%	44%	0%	0%	0.435	1	1	1
JP	5L value set				0.09 (0.05; 0.13)	39 (-416; 488)	418	56%	44%	0%	0%	0.435	1	1	1
	5L to 3L crosswalk				0.11 (0.09; 0.13)	39 (-416; 488)	356	56%	44%	0%	0%	0.435	1	1	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental QALYs IQ (95% CI)	Incremental costs € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	p_{CE}(0)	p_{CE}(20000)	p_{CE}(30000)	p_{CE}(50000)
NL	5L value set	(15)	**Moderate low back pain**	Large	0.19 (0.15; 0.24)	-303 (-916; 224)	-1591	14%	86%	0%	0%	0.858	1	1	1
	5L to 3L crosswalk				0.20 (0.17; 0.24)	-303 (-916; 224)	-1486	14%	86%	0%	0%	0.858	1	1	1
US	5L value set				0.22 (0.18; 0.27)	-303 (-916; 224)	-1368	14%	86%	0%	0%	0.858	1	1	1
	5L to 3L crosswalk				0.19 (0.16; 0.21)	-303 (-916; 224)	-1630	14%	86%	0%	0%	0.858	1	1	1
JP	5L value set				0.16 (0.12; 0.20)	-303 (-916; 224)	-1887	14%	86%	0%	0%	0.858	1	1	1
	5L to 3L crosswalk				0.19 (0.17; 0.21)	-303 (-916; 224)	-1640	14%	86%	0%	0%	0.858	1	1	1
NL	5L value set	(16)	**Severe low back pain**	Small	0.04 (0.001; 0.08)	178 (-285; 607)	4422	76%	22%	0%	2%	0.226	0.903	0.941	0.963
	5L to 3L crosswalk				0.03 (-0.002; 0.06)	178 (-285; 607)	5750	75%	22%	0%	3%	0.226	0.861	0.911	0.944
US	5L value set				0.03 (-0.01; 0.07)	178 (-285; 607)	6404	71%	21%	2%	6%	0.226	0.793	0.849	0.883
	5L to 3L crosswalk				0.02 (-0.004; 0.05)	178 (-285; 607)	8413	73%	22%	1%	4%	0.226	0.766	0.838	0.898
JP	5L value set				0.02 (-0.01; 0.06)	178 (-285; 607)	7616	71%	21%	2%	6%	0.226	0.761	0.831	0.874
	5L to 3L crosswalk				0.02 (-0.004; 0.04)	178 (-285; 607)	10327	73%	21%	2%	4%	0.226	0.714	0.806	0.875
NL	5L value set	(17)	**Severe low back pain**	Medium	0.11 (0.07; 0.15)	479 (-13; 939)	4338	97%	3%	0%	0%	0.025	0.999	1	1
	5L to 3L crosswalk				0.09 (0.06; 0.13)	479 (-13; 939)	5062	97%	3%	0%	0%	0.025	0.999	1	1
US	5L value set				0.10 (0.07; 0.14)	479 (-13; 939)	4701	97%	3%	0%	0%	0.025	0.998	0.999	1
	5L to 3L crosswalk				0.07 (0.05; 0.10)	479 (-13; 939)	6442	97%	3%	0%	0%	0.025	0.995	0.999	1
JP	5L value set				0.08 (0.05; 0.11)	479 (-13; 939)	6306	97%	3%	0%	0%	0.025	0.992	0.999	1
	5L to 3L crosswalk				0.06 (0.04; 0.08)	479 (-13; 939)	7621	97%	3%	0%	0%	0.025	0.985	0.999	1
NL	5L value set	(18)	**Severe low back pain**	Large	0.16 (0.13; 0.20)	249 (-161; 748)	1525	87%	13%	0%	0%	0.125	1	1	1
	5L to 3L crosswalk				0.15 (0.12; 0.18)	249 (-161; 748)	1693	87%	13%	0%	0%	0.125	1	1	1
US	5L value set				0.16 (0.12; 0.20)	249 (-161; 748)	1554	87%	13%	0%	0%	0.125	1	1	1
	5L to 3L crosswalk				0.12 (0.10; 0.15)	249 (-161; 748)	2040	87%	13%	0%	0%	0.125	1	1	1
JP	5L value set				0.12 (0.09; 0.15)	249 (-161; 748)	2079	87%	13%	0%	0%	0.125	1	1	1
	5L to 3L crosswalk				0.11 (0.09; 0.13)	249 (-161; 748)	2305	87%	13%	0%	0%	0.125	1	1	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental QALYs IQ (95% CI)	Incremental costs € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	5L value set	(19)	**Mild osteo-arthritis**	Small	0.03 (-0.01; 0.06)	82 (-401; 623)	2900	58%	36%	1%	5%	0.372	0.850	0.892	0.916
	5L to 3L crosswalk				0.03 (-0.01; 0.06)	82 (-401; 623)	2719	60%	36%	1%	3%	0.372	0.883	0.924	0.943
US	5L value set				0.04 (0.001; 0.08)	82 (-401; 623)	1945	61%	37%	1%	1%	0.372	0.937	0.957	0.971
	5L to 3L crosswalk				0.03 (0.001; 0.05)	82 (-401; 623)	2925	61%	37%	1%	1%	0.372	0.893	0.941	0.963
JP	5L value set				0.02 (-0.01; 0.05)	82 (-401; 623)	3529	58%	35%	2%	5%	0.372	0.821	0.861	0.896
	5L to 3L crosswalk				0.02 (-0.002; 0.04)	82 (-401; 623)	4173	60%	36%	1%	3%	0.372	0.820	0.879	0.918
NL	5L value set	(20)	**Mild osteo-arthritis**	Medium	0.10 (0.06; 0.13)	-125 (-629; 390)	-1299	32%	68%	0%	0%	0.677	1	1	1
	5L to 3L crosswalk				0.11 (0.08; 0.14)	-125 (-629; 390)	-1156	32%	68%	0%	0%	0.677	1	1	1
US	5L value set				0.10 (0.07; 0.14)	-125 (-629; 390)	-1198	32%	68%	0%	0%	0.677	1	1	1
	5L to 3L crosswalk				0.09 (0.06; 0.11)	-125 (-629; 390)	-1406	32%	68%	0%	0%	0.677	1	1	1
JP	5L value set				0.07 (0.04; 0.10)	-125 (-629; 390)	-1798	32%	68%	0%	0%	0.677	1	1	1
	5L to 3L crosswalk				0.07 (0.05; 0.09)	-125 (-629; 390)	-1781	32%	68%	0%	0%	0.677	1	1	1
NL	5L value set	(21)	**Mild osteo-arthritis**	Large	0.14 (0.12; 0.17)	-270 (-708; 179)	-1883	11%	89%	0%	0%	0.887	1	1	1
	5L to 3L crosswalk				0.16 (0.13; 0.18)	-270 (-708; 179)	-1741	11%	89%	0%	0%	0.887	1	1	1
US	5L value set				0.15 (0.11; 0.18)	-270 (-708; 179)	-1853	11%	89%	0%	0%	0.887	1	1	1
	5L to 3L crosswalk				0.13 (0.11; 0.15)	-270 (-708; 179)	-2115	11%	89%	0%	0%	0.887	1	1	1
JP	5L value set				0.11 (0.08; 0.13)	-270 (-708; 179)	-2574	11%	89%	0%	0%	0.887	1	1	1
	5L to 3L crosswalk				0.11 (0.09; 0.13)	-270 (-708; 179)	-2467	11%	89%	0%	0%	0.887	1	1	1
NL	5L value set	(22)	**Moderate osteoarthritis**	Small	0.03 (-0.01; 0.06)	-175 (-604; 296)	-6847	21%	72%	5%	2%	0.768	0.946	0.950	0.946
	5L to 3L crosswalk				0.02 (-0.01; 0.05)	-175 (-604; 296)	-8488	21%	69%	8%	2%	0.768	0.919	0.921	0.918
US	5L value set				0.03 (-0.01; 0.06)	-175 (-604; 296)	-6160	21%	73%	4%	2%	0.768	0.953	0.954	0.955
	5L to 3L crosswalk				0.01 (-0.01; 0.04)	-175 (-604; 296)	-12052	20%	67%	10%	3%	0.768	0.893	0.898	0.895
JP	5L value set				0.03 (-0.001; 0.05)	-175 (-604; 296)	-6939	22%	75%	2%	1%	0.768	0.965	0.972	0.974
	5L to 3L crosswalk				0.01 (-0.01; 0.04)	-175 (-604; 296)	-12450	20%	69%	8%	3%	0.768	0.906	0.915	0.914

Country	Scoring method	Scenario	Patient population	Effect size	Incremental QALYs IQ (95% CI)	Incremental costs € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	5L value set	(23)	**Moderate osteoarthritis**	Medium	0.08 (0.05; 0.11)	11 (-484; 634)	134	49%	51%	0%	0%	0.506	1	1	1
	5L to 3L crosswalk				0.09 (0.06; 0.12)	11 (-484; 634)	120	49%	51%	0%	0%	0.506	1	1	1
US	5L value set				0.09 (0.05; 0.12)	11 (-484; 634)	125	49%	51%	0%	0%	0.506	1	1	1
	5L to 3L crosswalk				0.07 (0.05; 0.09)	11 (-484; 634)	154	49%	51%	0%	0%	0.506	1	1	1
JP	5L value set				0.06 (0.03; 0.08)	11 (-484; 634)	180	49%	51%	0%	0%	0.506	0.998	1	1
	5L to 3L crosswalk				0.05 (0.03; 0.07)	11 (-484; 634)	203	49%	51%	0%	0%	0.506	0.999	1	1
NL	5L value set	(24)	**Moderate osteoarthritis**	Large	0.13 (0.10; 0.15)	-269 (-781; 259)	-2144	16%	84%	0%	0%	0.836	1	1	1
	5L to 3L crosswalk				0.14 (0.11; 0.17)	-269 (-781; 259)	-1938	16%	84%	0%	0%	0.836	1	1	1
US	5L value set				0.14 (0.11; 0.17)	-269 (-781; 259)	-1965	16%	84%	0%	0%	0.836	1	1	1
	5L to 3L crosswalk				0.11 (0.09; 0.14)	-269 (-781; 259)	-2355	16%	84%	0%	0%	0.836	1	1	1
JP	5L value set				0.10 (0.07; 0.12)	-269 (-781; 259)	-2731	16%	84%	0%	0%	0.836	1	1	1
	5L to 3L crosswalk				0.09 (0.07; 0.11)	-269 (-781; 259)	-2896	16%	84%	0%	0%	0.836	1	1	1
NL	5L value set	(25)	**Severe osteoarthritis**	Small	0.04 (-0.01; 0.09)	386 (-89; 893)	10222	87%	6%	0%	7%	0.063	0.712	0.798	0.862
	5L to 3L crosswalk				0.03 (-0.01; 0.08)	386 (-89; 893)	11157	87%	6%	0%	7%	0.063	0.696	0.799	0.861
US	5L value set				0.05 (-0.004; 0.11)	386 (-89; 893)	7693	90%	6%	0%	4%	0.063	**0.820**	0.890	0.929
	5L to 3L crosswalk				0.04 (-0.001; 0.08)	386 (-89; 893)	10566	91%	6%	0%	3%	0.063	**0.746**	0.861	0.927
JP	5L value set				0.04 (-0.01; 0.09)	386 (-89; 893)	10189	88%	6%	0%	6%	0.063	0.727	0.815	0.879
	5L to 3L crosswalk				0.03 (0.005; 0.07)	386 (-89; 893)	11441	92%	6%	0%	2%	0.063	0.751	0.874	0.939
NL	5L value set	(26)	**Severe osteoarthritis**	Medium	0.14 (0.09; 0.18)	-171 (-645; 318)	-1228	23%	77%	0%	0%	0.772	1	1	1
	5L to 3L crosswalk				0.15 (0.10; 0.19)	-171 (-645; 318)	-1177	23%	77%	0%	0%	0.772	1	1	1
US	5L value set				0.15 (0.10; 0.20)	-171 (-645; 318)	-1135	23%	77%	0%	0%	0.772	1	1	1
	5L to 3L crosswalk				0.12 (0.09; 0.16)	-171 (-645; 318)	-1386	23%	77%	0%	0%	0.772	1	1	1
JP	5L value set				0.11 (0.06; 0.16)	-171 (-645; 318)	-1561	23%	77%	0%	0%	0.772	1	1	1
	5L to 3L crosswalk				0.10 (0.07; 0.13)	-171 (-645; 318)	-1658	23%	77%	0%	0%	0.772	1	1	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental QALYs IQ (95% CI)	Incremental costs € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	5L value set	(27)	**Severe osteoarthritis**	Large	0.21 (0.16; 0.25)	-162 (-631; 297)	-781	25%	75%	0%	0%	0.751	1	1	1
	5L to 3L crosswalk				0.22 (0.18; 0.25)	-162 (-631; 297)	-747	25%	75%	0%	0%	0.751	1	1	1
US	5L value set				0.22 (0.17; 0.27)	-162 (-631; 297)	-746	25%	75%	0%	0%	0.751	1	1	1
	5L to 3L crosswalk				0.19 (0.16; 0.22)	-162 (-631; 297)	-844	25%	75%	0%	0%	0.751	1	1	1
JP	5L value set				0.17 (0.13; 0.22)	-162 (-631; 297)	-930	25%	75%	0%	0%	0.751	1	1	1
	5L to 3L crosswalk				0.18 (0.16; 0.21)	-162 (-631; 297)	-894	25%	75%	0%	0%	0.751	1	1	1
NL	5L value set	(28)	**Mild cancer**	Small	0.02 (-0.02; 0.05)	185 (-143; 726)	16790	74%	9%	1%	16%	0.104	0.564	0.653	0.735
	5L to 3L crosswalk				0.01 (-0.02; 0.05)	185 (-143; 726)	21052	69%	9%	1%	21%	0.104	0.510	0.600	0.668
US	5L value set				0.02 (-0.02; 0.06)	185 (-143; 726)	14746	76%	10%	1%	13%	0.104	0.604	0.698	**0.765**
	5L to 3L crosswalk				0.01 (-0.02; 0.04)	185 (-143; 726)	29126	67%	9%	2%	22%	0.104	0.426	0.523	**0.618**
JP	5L value set				0.02 (-0.01; 0.05)	185 (-143; 726)	15272	80%	10%	0%	10%	0.104	0.605	**0.719**	0.807
	5L to 3L crosswalk				0.02 (-0.004; 0.05)	185 (-143; 726)	11749	86%	10%	0%	39%	0.104	0.722	**0.830**	0.902
NL	5L value set	(29)	**Mild cancer**	Medium	0.09 (0.05; 0.12)	-182 (-634; 316)	-2049	21%	79%	0%	0%	0.787	1	1	1
	5L to 3L crosswalk				0.08 (0.05; 0.12)	-182 (-634; 316)	-2192	21%	79%	0%	0%	0.787	1	1	1
US	5L value set				0.09 (0.06; 0.13)	-182 (-634; 316)	-1921	21%	79%	0%	0%	0.787	1	1	1
	5L to 3L crosswalk				0.07 (0.04; 0.09)	-182 (-634; 316)	-2748	21%	79%	0%	0%	0.787	1	1	1
JP	5L value set				0.07 (0.04; 0.10)	-182 (-634; 316)	-2666	21%	79%	0%	0%	0.787	1	1	1
	5L to 3L crosswalk				0.07 (0.04; 0.09)	-182 (-634; 316)	-2732	21%	79%	0%	0%	0.787	1	1	1
NL	5L value set	(30)	**Mild cancer**	Large	0.15 (0.12; 0.18)	29 (-492; 538)	189	55%	45%	0%	0%	0.455	1	1	1
	5L to 3L crosswalk				0.15 (0.12; 0.18)	29 (-492; 538)	194	55%	45%	0%	0%	0.455	1	1	1
US	5L value set				0.15 (0.12; 0.19)	29 (-492; 538)	187	55%	45%	0%	0%	0.455	1	1	1
	5L to 3L crosswalk				0.11 (0.09; 0.14)	29 (-492; 538)	254	55%	45%	0%	0%	0.455	1	1	1
JP	5L value set				0.12 (0.09; 0.15)	29 (-492; 538)	246	55%	45%	0%	0%	0.455	1	1	1
	5L to 3L crosswalk				0.12 (0.09; 0.14)	29 (-492; 538)	247	55%	45%	0%	0%	0.455	1	1	1

Country	Scoring method	Scenario	Patient population	Effect size	Incremental QALYs IQ (95% CI)	Incremental costs € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	5L value set	(31)	**Moderate cancer**	Small	0.05 (-0.001; 0.11)	345 (-190; 916)	6643	86%	10%	1%	3%	0.102	0.847	0.903	0.928
	5L to 3L crosswalk				0.07 (0.02; 0.12)	345 (-190; 916)	4897	89%	10%	0%	1%	0.102	0.955	0.980	0.990
US	5L value set				0.06 (0.005; 0.12)	345 (-190; 916)	5689	87%	10%	1%	2%	0.102	0.888	0.925	0.948
	5L to 3L crosswalk				0.05 (0.01; 0.09)	345 (-190; 916)	7488	88%	10%	0%	2%	0.102	0.855	0.928	0.958
JP	5L value set				0.06 (0.01; 0.12)	345 (-190; 916)	5764	88%	10%	1%	1%	0.102	**0.895**	**0.936**	**0.960**
	5L to 3L crosswalk				0.005 (-0.03; 0.04)	345 (-190; 916)	72291	52%	6%	4%	38%	0.102	**0.270**	**0.349**	**0.425**
NL	5L value set	(32)	**Moderate cancer**	Medium	0.11 (0.07; 0.14)	-260 (-696; 178)	-2423	13%	87%	0%	0%	0.869	1	1	1
	5L to 3L crosswalk				0.12 (0.09; 0.15)	-260 (-696; 178)	-2129	13%	87%	0%	0%	0.869	1	1	1
US	5L value set				0.12 (0.08; 0.15)	-260 (-696; 178)	-2213	13%	87%	0%	0%	0.869	1	1	1
	5L to 3L crosswalk				0.11 (0.08; 0.13)	-260 (-696; 178)	-2433	13%	87%	0%	0%	0.869	1	1	1
JP	5L value set				0.08 (0.04; 0.11)	-260 (-696; 178)	-3217	13%	87%	0%	0%	0.869	1	1	1
	5L to 3L crosswalk				0.11 (0.09; 0.13)	-260 (-696; 178)	-2358	13%	87%	0%	0%	0.869	1	1	1
NL	5L value set	(33)	**Moderate cancer**	Large	0.30 (0.26; 0.34)	-236 (-724; 197)	-778	15%	85%	0%	0%	0.848	1	1	1
	5L to 3L crosswalk				0.27 (0.24; 0.30)	-236 (-724; 197)	-874	15%	85%	0%	0%	0.848	1	1	1
US	5L value set				0.27 (0.22; 0.31)	-236 (-724; 197)	-889	15%	85%	0%	0%	0.848	1	1	1
	5L to 3L crosswalk				0.19 (0.17; 0.21)	-236 (-724; 197)	-1243	15%	85%	0%	0%	0.848	1	1	1
JP	5L value set				0.14 (0.10; 0.18)	-236 (-724; 197)	-1682	15%	85%	0%	0%	0.848	1	1	1
	5L to 3L crosswalk				0.14 (0.12; 0.17)	-236 (-724; 197)	-1665	15%	85%	0%	0%	0.848	1	1	1
NL	5L value set	(34)	**Severe cancer**	Small	0.04 (-0.02; 0.10)	307 (-190; 832)	7474	79%	11%	1%	9%	0.118	0.772	0.824	0.860
	5L to 3L crosswalk				0.03 (-0.02; 0.09)	307 (-190; 832)	9040	78%	11%	1%	10%	0.118	0.722	0.786	0.835
US	5L value set				0.05 (-0.01; 0.12)	307 (-190; 832)	6039	82%	12%	0%	6%	0.118	**0.838**	**0.877**	0.906
	5L to 3L crosswalk				0.02 (-0.01; 0.07)	307 (-190; 832)	12906	76%	11%	1%	12%	0.118	**0.623**	**0.713**	0.784
JP	5L value set				0.04 (-0.01; 0.11)	307 (-190; 832)	6839	81%	12%	0%	7%	0.118	**0.815**	0.853	0.887
	5L to 3L crosswalk				0.03 (-0.01; 0.06)	307 (-190; 832)	11741	80%	11%	0%	8%	0.118	**0.674**	0.768	0.846

Country	Scoring method	Scenario	Patient population	Effect size	Incremental QALYs IQ (95% CI)	Incremental costs € (95% CI)	ICER €/point	Distribution CE-plane (%)				Probability of cost-effectiveness			
								NE	SE	SW	NW	$p_{CE}(0)$	$p_{CE}(20000)$	$p_{CE}(30000)$	$p_{CE}(50000)$
NL	5L value set	(35)	**Severe cancer**	Medium	0.18 (0.12; 0.24)	386 (-38; 854)	2171	95%	5%	0%	0%	0.045	1	1	1
	5L to 3L crosswalk				0.17 (0.12; 0.22)	386 (-38; 854)	2214	95%	5%	0%	0%	0.045	1	1	1
US	5L value set				0.19 (0.13; 0.25)	386 (-38; 854)	2032	95%	5%	0%	0%	0.045	1	1	1
	5L to 3L crosswalk				0.14 (0.1; 0.18)	386 (-38; 854)	2711	95%	5%	0%	0%	0.045	1	1	1
JP	5L value set				0.14 (0.08; 0.2)	386 (-38; 854)	2686	95%	5%	0%	0%	0.045	1	1	1
	5L to 3L crosswalk				0.12 (0.09; 0.16)	386 (-38; 854)	3173	95%	5%	0%	0%	0.045	1	1	1
NL	5L value set	(36)	**Severe cancer**	Large	0.27 (0.21; 0.33)	-162 (-640; 305)	-595	24%	76%	0%	0%	0.756	1	1	1
	5L to 3L crosswalk				0.29 (0.24; 0.34)	-162 (-640; 305)	-551	24%	76%	0%	0%	0.756	1	1	1
US	5L value set				0.30 (0.24; 0.36)	-162 (-640; 305)	-5360.	24%	76%	0%	0%	0.756	1	1	1
	5L to 3L crosswalk				0.27 (0.23; 0.31)	-162 (-640; 305)	-604	24%	76%	0%	0%	0.756	1	1	1
JP	5L value set				0.24 (0.18; 0.30)	-162 (-640; 305)	-681	24%	76%	0%	0%	0.756	1	1	1
	5L to 3L crosswalk				0.25 (0.21; 0.29)	-162 (-640; 305)	-647	24%	76%	0%	0%	0.756	1	1	1

NL: the Netherlands. US: United States. JP: Japan. QALY: quality-adjusted life-year. IQ: incremental QALY. 95% CI: 95% confidence interval. ICER: Incremental Cost-Effectiveness Ratio. NE: northeast. SE: southeast. SW: southwest. NW: northwest. p_{CE}: probability of the intervention being cost-effective compared to control.

CHAPTER 4

Assessing the impact of EQ-5D country-specific value sets on cost-utility outcomes

Johanna M. van Dongen*, Ângela Jornada Ben*, Aureliano P. Finch,
Milou M. M. Rossenaar, Karolien E. M. Biesheuvel-Leliefeld, Adrie T. Apeldoorn,
Maurits W. van Tulder, Harm van Marwijk, Judith E. Bosmans
* Contributed equally

Medical Care, 2021; 59(1): 82 – 90.

Abstract

Purpose

To assess the impact of EQ-5D country-specific value sets on cost-utility outcomes.

Methods

Data from 2 randomized controlled trials on low back pain (LBP) and depression were used. 3L value sets were identified from the EuroQol Web site. A nonparametric crosswalk was employed for each tariff to obtain the likely 5L values. Differences in quality-adjusted life years (QALYs) between countries were tested using paired t tests, with United Kingdom as reference. Cost-utility outcomes were estimated for both studies and both EQ-5D versions, including differences in QALYs and cost-effectiveness acceptability curves.

Results

For the 3L, QALYs ranged between 0.650 (Taiwan) and 0.892 (United States) in the LBP study and between 0.619 (Taiwan) and 0.879 (United States) in the depression study. In both studies, most country-specific QALY estimates differed statistically significantly from that of the United Kingdom. Incremental cost-effectiveness ratios ranged between €2044/QALY (Taiwan) and €5897/QALY (Zimbabwe) in the LBP study and between €38,287/QALY (Singapore) and €96,550/QALY (Japan) in the depression study. At the NICE threshold of €23,300/QALY (≈£20,000/QALY), the intervention's probability of being cost-effective versus control ranged between 0.751 (Zimbabwe) and 0.952 (Taiwan) and between 0.230 (Canada) and 0.396 (Singapore) in the LBP study and depression study, respectively. Similar results were found for the 5L, with extensive differences in ICERs and moderate differences in the probability of cost-effectiveness.

Conclusions

This study indicates that the use of different EQ-5D country-specific value sets impacts on cost-utility outcomes. Therefore, to account for the fact that health state preferences are affected by sociocultural differences, relevant country-specific value sets should be used.

Introduction

In economic evaluations of health care interventions, quality-adjusted life years (QALYs) are often used as a metric of health effects. QALYs combine the length of life and the health-related quality of life (HRQoL) into 1 metric. HRQoL is measured in terms of utility units. A utility is a weight that typically indicates the general public's strength of preference or desirability for a given health state or condition. A utility of zero represents dead and a utility of 1 full health, whereas negative values are assigned to those states considered worse than death.[1,2]

Utility values are typically estimated using generic preference-based measures of health,[3,4] including the EuroQol 5-dimension questionnaire (EQ-5D),[5] the short form 36 health survey questionnaire (SF-6D),[6] the Health Utilities Index Mark 3 (HUI3),[7] the Assessment of Quality of Life instrument (AQoL),[8] and the 15-dimensional measure of HRQoL (15D).[9] These measures describe health in terms of dimensions and severity levels, and come with a value set or tariff that assigns a utility value to each of the health states described.[2] Such value sets have been traditionally estimated using cardinal response methods,[10] such as the time trade-off (TTO), the visual analogue scale (VAS), and the standard gamble (SG), and more recently using ordinal response methods, such as a Discrete Choice Experiment including duration (DCE_{TTO}) on a subset of all possible health states.[11,12] The derived data are modelled and result in a tariff through which the full set of health states can be obtained.[2,13,14]

The EQ-5D[5] is, among the other generic preference-based measures, the most commonly used worldwide.[2] The original EQ-5D uses a classification system consisting of 5 health dimensions, that is mobility, self-care, usual activities, pain/discomfort, and anxiety/depression, and 3 severity levels, that is no problems, some problems, severe problems (further referred to as the 3L).[5] To increase the 3L's sensitivity to changes in health and to reduce its commonly observed ceiling effects,[15] a 5-level version of the EQ-5D was developed (further referred to as the 5L).[16] The 5L maintained the same health dimensions, but described health using 5 severity levels. Currently, a large number of value sets is available for estimating EQ-5D health states' utility values. As evidence suggests that health state preferences are affected by sociocultural differences, such value sets are typically developed per country separately.[17]

Previous studies indicate that different country-specific EQ-5D value sets result in different utility values and QALY estimates.[13,18,19] However, this does not necessarily mean that the identified differences also impact cost-utility outcomes, and thus, the health technology assessment process. This is because the identified differences across countries may be equal in the intervention and control group, thereby not affecting cost-utility outcomes, such as incremental cost-effectiveness ratios (ICERs) and cost-effectiveness acceptability curves (CEAC).[19] To address this research gap, the present study aimed to explore the impact of country-specific EQ-5D value sets on cost-utility outcomes.

Methods

The impact of country-specific EQ-5D value sets on cost-utility outcomes was explored by comparing the results of 2 cost-utility analyses, employing different EQ-5D country-specific value sets, for both the 3L and 5L.

Data of 2 published randomized controlled trials performed in the Netherlands were used to estimate cost-utility outcomes.[20,21] As the EQ-5D focuses on physical as well as mental aspects of health, one of these studies included patients with a physical disorder [i.e., low back pain (LBP)] and the other included patients with a mental disorder (i.e., depression).

The LBP study evaluated the cost-effectiveness of a treatment-based classification system for subacute and chronic LBP patients in comparison with usual care. The study had a 52-week follow-up. A total of 156 patients were included, of which 74 in the intervention group and 82 in the control group.[20] The depression study evaluated a nurse-led self-help treatment in combination with usual care for recurrent depression patients in comparison with usual care alone. The study had a 65-week follow-up. For the purpose of this study, only data collected up until 52 weeks were used. A total of 248 patients were included, of which 124 in the intervention group and 124 in the control group.[21]

In both studies, costs were measured from a societal perspective, including health care and lost productivity costs. All costs were expressed in Euros 2013. A detailed description of the example studies' measurement and valuation of costs can be found elsewhere.[20–23] Baseline characteristics of the example studies' participants can be found in Appendix 1, Supplemental Digital Content 1 (https://links.lww.com/MLR/C111). A summary of the health states of the example studies' participants can be found in Appendix 2, Supplemental Digital Content 2 (https://links.lww.com/MLR/C112).

Country-specific EQ-5D utility values

In the LBP study, the 3L was administered at baseline, 8, 26, and 52 weeks. In the depression study, it was administered at baseline, 13, 26, 39, and 52 weeks.

Country-specific 3L value sets were identified from the EuroQol Web site (http://euroqol.org). In total, 33 3L value sets were available on the EuroQol Web site, of which 23 were developed using TTO and 10 were developed using VAS or a combination of both. Of them, we only included TTO-based value sets, because previous evidence indicates that TTO values tend to be systematically higher than VAS values, irrespective of country.[24] Herewith, we wanted to ensure that possible differences across countries were due to sociocultural differences, rather than the previously established differences between VAS and TTO values. On top of that, value sets were excluded if they were derived from populations other than the general population, and if they contained interaction terms, quadratic variables, and/or additional health states, such as death or unconscious. Eventually, 16 different country-specific value sets were included (Appendix 3, Supplemental Digital Content 3, https://links.lww.com/MLR/C113).[18,25–46] The participants' EQ-5D health states were converted into 3L utility values using all of the identified value sets. In addition, 5L utility values were estimated for all countries using the crosswalk approach.[47] To do that, the observed 3L health states were converted into possible 5L health states using the nonparametric crosswalk developed by van Hout et al.[47]

Statistical analysis

Missing EQ-5D descriptive system data and cost data were imputed using multivariate imputation by chained equations.48 Pooled estimates were calculated using Rubin's rules.[48] Imputation models were constructed per study, and included total costs, EQ-5D item responses at all measurement points, variables related to the "missingness" of data, variables differing between treatment groups at baseline, and variables that were related to the outcomes. Imputation models were stratified for study group.

Cost-utility outcomes were estimated for both studies, all countries, and both versions of the EQ-5D. QALYs were estimated by multiplying the participants' utility value of a health state by the duration of time they spent in that health state using linear interpolation between measurement points. Subsequently, it was explored whether there were minimally clinically important differences between the country-specific QALY estimates and the UK QALY estimate. The UK QALY estimate was used as reference, as a quick Google Scholar search indicated the UK value set to be the most frequently referenced one. The minimally clinically important difference for QALYs was set at 0.057.49,50 Paired t tests were used to test whether country-specific QALY estimates statistically significantly differed from that of the United Kingdom.

ICERs were calculated by dividing the mean differences in costs by the mean differences in QALYs. Mean differences in costs and QALYs across groups were estimated using seemingly unrelated regression analyses.[51] For all country-specific ICERs, it was explored whether they were above or below the upper and lower bound of the Dutch willingness-to-pay threshold (ie, €10,000/QALY to €80,000/QALY) as well as the NICE threshold of £20,000/QALY (ie, about €23,300/QALY). Bias-corrected and accelerated bootstrapping with 5000 replications was used to estimate the uncertainty surrounding cost differences and ICERs. Bootstrapped cost-effect pairs were plotted on cost-effectiveness planes. To provide a summary measure of the joint uncertainty surrounding cost and QALY differences, CEAC were constructed.[52] A CEAC provides insight into the probability of an intervention being cost-effective compared with a control for a range of ceiling ratios (ie, the maximum amount of money decision-makers are willing to pay per additional unit of effect).[52] Analyses were performed in STATA version 12 and statistical significance was set at $P<0.05$.

Sensitivity analysis

For countries in which an EQ-5D value set is already available for the 5L (see Appendix 1, Supplemental Digital Content 1, https://links.lww.com/MLR/C111), alternative 5L utility values were estimated. In these cases, the transition probabilities of van Hout et al.[47] were multiplied by the 5L utility values instead of the 3L utility values. Using the alternative 5L utility values, cost-utility outcomes were estimated for both studies and all applicable countries again.

Results

Quality-adjusted Life Years

In the LBP study, QALY estimates derived from the 3L ranged between 0.650 (Taiwan) and 0.892 (United States) (Table 1). Except for the Japanese case, country-specific QALY estimates statistically significantly differed from that of the United Kingdom. For Germany, the United States, South Korea, Singapore, and Taiwan, this difference was also clinically relevant (Table 1). Similar results were found when calculating QALYs for the 5L using the nonparametric crosswalk, with QALY estimates ranging between 0.644 (Taiwan) and 0.884 (United States); 14 of 15 country-specific QALY estimates statistically significantly differing from that of the United Kingdom, and 4 of 15 country-specific QALY estimates clinically relevantly differing from that of the United Kingdom.

In the depression study, QALY estimates derived from the 3L ranged between 0.619 (Taiwan) and 0.879 (United States) (Table 1). Once again, except for the Japanese case, country-specific QALY estimates statistically significantly differed from that of the United Kingdom, and for Germany, the United States, South Korea, Singapore, and Taiwan this difference was clinically relevant (Table 1). Similar results were found when calculating QALYs for the 5L using the nonparametric crosswalk, with these ranging between 0.612 (Taiwan) and 0.870 (United States), 14 of 15 country-specific QALY values statistically significantly differing from that of the United Kingdom, and 4 of 15 country-specific QALY values clinically relevantly differing from that of the United Kingdom.

Incremental Cost-Effectiveness Ratio and Cost-Effectiveness Acceptability Curves

In the LBP study, ICERs derived from the 3L ranged between €2044 per QALY (Taiwan) and €5897 per QALY (Zimbabwe) (Table 2). All ICERs were below the Dutch as well as the NICE threshold. The corresponding probabilities of the intervention being cost-effective compared with the control differed moderately across countries; at the lower bound of the Dutch threshold (ie, €10,000 per QALY), this probability ranged between 0.569 (Zimbabwe) and 0.810 (Taiwan) and at the NICE threshold it ranged between 0.751 (Zimbabwe) and 0.952 (Taiwan) (Table 4; Figure 1A). Similar results were found for the 5L, with extensive differences in ICERs and the probability of cost-effectiveness ranging between 0.565 (Zimbabwe) and 0.806 (Taiwan) at the lower bound of the Dutch threshold (Tables 2–4).

In the depression study, ICERs derived from the 3L ranged between €38,287 per QALY (Singapore) and €96,550 per QALY (Japan) (Table 3). None of the ICERs was below the lower bound of the Dutch threshold and the NICE threshold, and 14 (88%) ICERs were below the upper bound of the Dutch threshold. Again, differences were found across countries in the probability of the intervention being cost-effective compared with the control. At the NICE threshold, this probability ranged between 0.230 (Canada) and 0.396 (Singapore) (Table 4; Figure 1B). Similar results were found for the 5L, with extensive differences in ICERs and the probability of cost-effectiveness ranging between 0.283 (Japan) and 0.394 (Singapore) at the NICE threshold (Tables 2–4).

Please note that countries with the most similar QALY estimates (Table 1; e.g., United Kingdom and Japan) do not necessarily have the most similar ICERs (Tables 2, 3). This is due to the fact that QALYs were estimated for all participants of the example studies, whereas ICERs were based on the mean differences in QALYs across study groups.

Sensitivity analysis

Results of the sensitivity analysis were similar to those of the main analysis, with QALY estimates and ICERs differing extensively across countries and moderate differences across countries in the probability of the intervention being cost-effective compared with the control at all thresholds, in both studies (Tables 2–4).

Table 1 | Mean QALYs after 52 weeks based on the 3L and 5L, differences between QALYs of the various countries and that of the United Kingdom – Based on the Low Back Pain study and the Depression study

	Low Back Pain study		Depression study	
	Mean QALY (95% CI)	QALY country – QALY UK (95% CI)	Mean QALY (95% CI)	QALY country – QALY UK (95% CI)
Country – 3L				
United Kingdom	0.767 (0.747; 0.786)	–	0.757 (0.739; 0.755)	–
Spain	0.809 (0.790; 0.828)	0.042 (0.039; 0.045)	0.802 (0.784; 0.819)	0.045 (0.040; 0.051)
Japan	0.772 (0.758; 0.787)	0.005 (-0.002; 0.013)	0.762 (0.750; 0.775)	0.005 (-0.003; 0.014)
Zimbabwe	0.788 (0.777;0.799)	0.021 (0.013; 0.030)	0.812 (0.800; 0.823)	0.055 (0.048; 0.062)
Germany	0.856 (0.841; 0.871)	0.089 (0.083; 0.095)	0.869 (0.854; 0.833)	0.112 (0.106; 0.117)
United States	0.892 (0.878; 0.906)	0.125 (0.118; 0.133)	0.879 0.866; 0.892)	0.122 (0.115; 0.129)
Netherlands	0.807 (0.790; 0.824)	0.040 (0.036; 0.043)	0.776 (0.760; 0.792)	0.019 (0.015; 0.022)
South Korea	0.867 (0.855; 0.879)	0.100 (0.091; 0.108)	0.858 (0.847; 0.868)	0.101 (0.092; 0.109)
Denmark	0.799 (0.783; 0.815)	0.032 (0.028; 0.036)	0.784 (0.770; 0.798)	0.027 (0.022; 0.031)
France	0.714 (0.691; 0.736)	-0.052 (-0.060; -0.046)	0.704 (0.684; 0.724)	-0.053 (-0.058; -0.048)
Thailand	0.712 (0.693; 0.732)	-0.053 (-0.060; -0.049)	0.709 (0.693; 0.726)	-0.048 (-0.053; -0.043)
Canada	0.812 (0.798; 0.826)	0.045 (0.039; 0.051)	0.797 (0.784; 0.809)	0.040 (0.033; 0.046)
China	0.815 (0.800; 0.830)	0.048 (0.042; 0.054)	0.805 (0.793; 0.818)	0.048 (0.041; 0.055)
Italy	0.804 (0.788; 0.820)	0.037 (0.032; 0.042)	0.791 (0.777; 0.805)	0.034 (0.029; 0.040)
Singapore	0.690 (0.663; 0.717)	-0.077 (-0.087; -0.066)	0.665 (0.640; 0.689)	-0.092 (-0.101; -0.083)
Taiwan	0.650 (0.621; 0.675)	-0.117 (-0.124; -0.109)	0.619 (0.597; 0.641)	-0.138 (-0.145; -0.132)
Country – 5L				
United Kingdom	0.759 (0.740; 0.779)	–	0.749 (0.730; 0.767)	–
Spain	0.801 (0.783; 0.820	0.042 (0.039; 0.045)	0.793 (0.775; 0.810)	0.044 (0.041; 0.047)
Japan	0.765 (0.751; 0.779)	0.006 (-0.002; 0.013)	0.754 (0.742; 0.766)	0.005 (-0.003; 0.014)
Zimbabwe	0.781 (0.770; 0.792)	0.022 (0.013; 0.030)	0.780 (0.770; 0.789)	0.031 (0.022; 0.040)
Germany	0.848 (0.833; 0.862)	0.089 (0.083; 0.094)	0.859 (0.845; 0.873)	0.110 (0.105; 0.115)
United States	0.884 (0.870; 0.897)	0.125 (0.117; 0.132)	0.870 (0.857; 0.882)	0.121 (0.114; 0.128)
Netherlands	0.799 (0.782; 0.816)	0.039 (0.036; 0.043)	0.767 (0.751; 0.783)	0.018 (0.015; 0.022)
South Korea	0.858 (0.847; 0.870)	0.099 (0.091; 0.107)	0.849 (0.839; 0.859)	0.100 (0.092; 0.108)
Denmark	0.791 (0.775; 0.807)	0.032 (0.027; 0.036)	0.775 (0.761; 0.789)	0.026 (0.021; 0.030)
France	0.843 (0.823; 0.864)	0.084 (0.076; 0.092)	0.809 (0.791; 0.826)	0.060 (0.052; 0.067)
Thailand	0.705 (0.686; 0.725)	-0.054 (-0.060; -0.048)	0.702 (0.685; 0.718)	-0.047 (-0.052; -0.042)

	Low Back Pain study		Depression study	
	Mean QALY (95% CI)	QALY country – QALY UK (95% CI)	Mean QALY (95% CI)	QALY country – QALY UK (95% CI)
Canada	0.804 (0.790; 0.818)	0.045 (0.038; 0.051)	0.788 (0.776; 0.800)	0.039 (0.033; 0.046)
China	0.807 (0.793; 0.822)	0.048 (0.042; 0.054)	0.797 (0.784; 0.809)	0.048 (0.041; 0.055)
Italy	0.796 (0.780; 0.812)	0.037 (0.0.32; 0.042)	0.783 (0.769; 0.797)	0.034 (0.029; 0.039)
Singapore	0.683 (0.656; 0.710)	-0.076 (-0.086; -0.066)	0.657 (0.633; 0.681)	-0.092 (-0.100; -0.082)
Taiwan	0.644 (0.619; 0.668)	-0.115 (-0.123; -0.108)	0.612 (0.590; 0.633)	-0.137 (-0.143; -0.130)

Table 2 | Differences in pooled mean costs and effects (95% confidence intervals), incremental cost-effectiveness ratios, and the distribution of incremental cost–effect pairs around the quadrants of the cost-effectiveness planes − based on the Low Back Pain study

	ΔC (95% CI) €	ΔE (95% CI) Points	ICER €/point	Distribution CE-plane (%)			
				NE[1]	SE[2]	SW[3]	NW[4]
Country – 3L							
United Kingdom	122 (-725; 1045)	0.047 (0.008; 0.085)	2,605	59.4	39.7	0.3	0.6
Spain	122 (-725; 1045)	0.042 (0.042; 0.079)	2,913	59.1	39.5	0.5	0.9
Japan	122 (-725; 1045)	0.036 (0.007; 0.064)	3,421	59.6	39.8	0.2	0.4
Zimbabwe	122 (-725; 1045)	0.021 (-0.001; 0.043)	5,897	57.9	38.8	1.2	2.1
Germany	122 (-725; 1045)	0.036 (0.006; 0.066)	3,385	59.4	39.7	0.3	0.6
United States	122 (-725; 1045)	0.025 (-0.003; 0.052)	4,945	57.5	38.6	1.4	2.5
Netherlands	122 (-725; 1045)	0.040 (0.007; 0.074)	3,022	59.4	39.7	0.3	0.6
South Korea	122 (-725; 1045)	0.025 (0.001; 0.048)	4,901	58.8	39.3	0.7	1.2
Denmark	122 (-725; 1045)	0.035 (0.003; 0.066)	3,510	59.0	39.4	0.6	1.0
France	122 (-725; 1045)	0.048 (0.003; 0.093)	2,526	58.9	39.4	0.6	1.1
Thailand	122 (-725; 1045)	0.049 (0.010; 0.087)	2,509	59.6	39.8	0.2	0.4
Canada	122 (-725; 1045)	0.031 (0.003; 0.059)	3,945	59.1	39.5	0.5	0.9
China	122 (-725; 1045)	0.030 (0.001;0.060)	4,034	58.6	39.2	0.8	1.4
Italy	122 (-725; 1045)	0.034 (0.002; 0.066)	3,575	58.9	39.3	0.7	1.1
Singapore	122 (-725; 1045)	0.053 (-0.0002; 0.106)	2,296	58.5	39.1	0.9	1.4
Taiwan	122 (-725; 1045)	0.060 (0.010; 0.109)	2,044	59.4	39.7	0.3	0.5
Country – 5L							
United Kingdom	122 (-725; 1045)	0.046 (0.008; 0.084)	2,647	59.4	39.7	0.3	0.6
Spain	122 (-725; 1045)	0.041 (0.004; 0.078)	2,963	59.0	39.5	0.5	1.0
Japan	122 (-725; 1045)	0.035 (0.007; 0.063)	3,487	59.6	39.8	0.2	0.4
Zimbabwe	122 (-725; 1045)	0.020 (-0.001; 0.042)	6,048	57.8	38.8	2	2
Germany	122 (-725; 1045)	0.035 (0.006; 0.065)	3,446	59.4	39.7	0.3	0.6
United States	122 (-725; 1045)	0.024 (-0.003; 0.051)	5,063	57.4	38.6	1.4	2.6
Netherlands	122 (-725; 1045)	0.040 (0.006; 0.073)	3,075	59.4	39.7	0.3	0.6
South Korea	122 (-725; 1045)	0.024 (0.001; 0.047)	5,020	58.8	39.3	0.7	1.2
Denmark	122 (-725; 1045)	0.034 (0.003; 0.065)	3,579	59.0	39.4	0.6	1.0
France	122 (-725; 1045)	0.038 (-0.002; 0.079)	3,182	58.0	38.8	1.2	2.0

	ΔC (95% CI) €	ΔE (95% CI) Points	ICER €/point	Distribution CE-plane (%)			
				NE[1]	SE[2]	SW[3]	NW[4]
Thailand	122 (-725; 1045)	0.048 (0.0100; 0.086)	2,548	59.6	39.8	0.2	0.4
Canada	122 (-725; 1045)	0.030 (0.003; 0.057)	4,035	59.1	39.5	0.5	0.9
China	122 (-725; 1045)	0.029 (0.001; 0.058)	4,121	58.6	39.2	0.8	1.4
Italy	122 (-725; 1045)	0.033 (0.002; 0.065)	3,644	58.9	39.3	0.7	1.1
Singapore	122 (-725; 1045)	0.052 (-0.0003; 0.105)	2,330	58.4	39.1	0.9	1.6
Taiwan	122 (-725; 1045)	0.059 (0.010; 0.108)	2,074	59.4	39.7	0.3	0.6
Sensitivity analysis							
United Kingdom	122 (-725; 1045)	0.075 (0.015; 0.134)	1,634	59.6	39.8	0.2	0.4
Spain	122 (-725; 1045)	0.076 (0.016; 0.137)	1,594	59.6	39.8	0.2	0.4
Japan	122 (-725; 1045)	0.02z3 (0.000; 0.046	5,262	58.6	39.1	0.9	1.4
Germany	122 (-725; 1045)	0.076 (0.015; 0.136)	1,609	59.6	39.8	0.2	0.4
Netherlands	122 (-725; 1045)	0.029 (-0.000; 0.058)	4,245	58.3	39.0	1.0	1.7
South Korea	122 (-725; 1045)	0.020 (0.001; 0.039)	6,045	59.9	39.4	0.6	0.1
Thailand	122 (-725; 1045)	0.072 (0.014; 0.130)	1,680	59.6	39.7	0.2	0.4

CI indicates confidence interval; CE, cost-effectiveness; ICER, incremental cost-effectiveness ratio.

Table 3 | Differences in pooled mean costs and effects (95% confidence intervals), incremental cost-effectiveness ratios, and the distribution of incremental cost-effect pairs around the quadrants of the cost-effectiveness planes – Based on the and Depression study

	ΔC (95% CI) €	ΔE (95% CI) Points	ICER €/point	Distribution CE-plane (%)			
				NE[1]	SE[2]	SW[3]	NW[4]
Country – 3L							
United Kingdom	1463 (-1734; 4322)	0.033 (-0.003; 0.069)	44,452	76.8	19.7	0.4	0.3
Spain	1463 (-1734; 4322)	0.032 (-0.002; 0.067)	45,366	77.1	19.8	0.3	2.8
Japan	1463 (-1734; 4322)	0.015 (-0.009; 0.039)	96,550	70.4	18.6	1.4	9.5
Zimbabwe	1463 (-1734; 4322)	0.017 (-0.005; 0.039)	87,779	73.7	19.3	0.8	6.2
Germany	1463 (-1734; 4322)	0.028 (0.000; 0.057)	51,715	77.5	19.9	0.2	2.4
United States	1463 (-1734; 4322)	0.026 (0.000; 0.051)	56,879	78.0	19.9	0.2	1.9
Netherlands	1463 (-1734; 4322)	0.029 (-0.003; 0.061)	50,574	76.7	19.7	0.4	3.3
South Korea	1463 (-1734; 4322)	0.019 (-0.025; 0.040)	78,462	76.4	19.7	0.4	3.5
Denmark	1463 (-1734; 4322)	0.024 (-0.004; 0.052)	61,266	76.0	19.6	0.5	3.9
France	1463 (-1734; 4322)	0.030 (-0.009; 0.069)	48,834	74.9	18.4	0.7	6.1
Thailand	1463 (-1734; 4322)	0.022 (-0.011; 0.054)	67,799	71.2	18.9	1.2	8.8
Canada	1463 (-1734; 4322)	0.019 (-0.006; 0.043)	78,345	74.1	19.3	0.7	5.8
China	1463 (-1734; 4322)	0.020 (-0.005; 0.045)	72,438	75.0	19.6	0.6	4.9
Italy	1463 (-1734; 4322)	0.024 (-0.004; 0.052)	61,384	75.7	19.7	0.5	4.1
Singapore	1463 (-1734; 4322)	0.038 (-0.010; 0.087)	38,287	74.4	19.5	0.6	5.4
Taiwan	1463 (-1734; 4322)	0.031 (-0.012; 0.074)	47,179	72.8	19.3	0.9	7.0

	ΔC (95% CI) €	ΔE (95% CI) Points	ICER €/point	Distribution CE-plane (%)			
				NE[1]	SE[2]	SW[3]	NW[4]
Country – 5L							
United Kingdom	1463 (-1750; 4368)	0.032 (-0.002 to 0.069)	45,050	76.8	19.9	0.3	3.0
Spain	1463 (-1750; 4368)	0.032 (-0.002 to 0.066)	46,013	77.0	19.9	0.3	2.8
Japan	1463 (-1750; 4368)	0.015 (-0.009 to 0.039)	98,726	70.1	18.8	1.4	9.8
Zimbabwe	1463 (-1750; 4368)	0.017 (-0.001 to 0.035)	86,867	76.9	19.9	0.3	2.9
Germany	1463 (-1750; 4368)	0.028 (-0.001 to 0.056)	52,593	77.6	22.0	0.2	2.3
United States	1463 (-1750; 4368)	0.025 (0.000 to 0.050)	57,827	78.0	20.0	0.2	1.8
Netherlands	1463 (-1750; 4368)	0.028 (-0.003 to 0.060)	51,298	76.6	19.8	0.3	3.2
South Korea	1463 (-1750; 4368)	0.018 (-0.028 to 0.038)	82,973	76.1	19.7	0.5	3.7
Denmark	1463 (-1750; 4368)	0.024 (-0.004 to 0.052)	60,914	76.1	19.7	0.5	3.7
France	1463 (-1750; 4368)	0.022 (-0.014 to 0.057)	67,444	69.7	18.9	1.2	10.2
Thailand	1463 (-1750; 4368)	0.021 (-0.011 to 0.054)	68,894	70.9	19.0	1.2	8.9
Canada	1463 (-1750; 4368)	0.018 (-0.006 to 0.042)	79,843	73.9	19.4	0.8	5.9
China	1463 (-1750; 4368)	0.020 (-0.004; 0.044)	73,847	74.9	19.6	0.6	4.9
Italy	1463 (-1750; 4368)	0.023 (-0.004; 0.051)	62,423	75.7	19.7	0.5	4.1
Singapore	1463 (-1750; 4368)	0.038 (-0.010; 0.086)	38,724	74.4	19.5	0.6	5.4
Taiwan	1463 (-1750; 4368)	0.031 (0.012; 0.073)	47,733	72.8	19.3	0.9	7.0
Sensitivity analysis							
United Kingdom	1463 (-1750; 4368)	0.018 (-0.031; 0.067)	82,226	59.3	16.7	3.5	20.6
Spain	1463 (-1750; 4368)	0.018 (-0.031; 0.068)	79,684	59.6	16.7	3.4	20.3
Japan	1463 (-1750; 4368)	0.013 (-0.006; 0.032)	109,920	71.9	19.1	1.1	7.9
Germany	1463 (-1750; 4368)	0.020 (-0.029; 0.070)	71,229	61.8	17.3	2.9	18.0
Netherlands	1463 (-1750; 4368)	0.023 (-0.003; 0.048)	64,703	76.2	19.7	0.4	3.6
South Korea	1463 (-1750; 4368)	0.012 (-0.004; 0.027)	125,958	73.5	19.3	0.8	6.4
Thailand	1463 (-1750; 4368)	0.017 (-0.031; 0.065)	85,764	58.8	16.6	3.6	21.0

CI indicates confidence inteval; CE, cost-effectiveness; ICER, incremental cost-effectiveness ratio.

Table 4 | Probabilities of cost-effectiveness at a ceiling ratio of €0 per QALY, €10,000 per QALY, €23,330 per QALY and €80,000 per QALY – Based on the Low Back Pain study and the Depression study

	Low Back Pain study – ceiling ratio				Depression study – ceiling ratio			
	€0/ QALY	€10,000/ QALY	€23,330/ QALY	€80,000/ QALY	€0/ QALY	€10,000/ QALY	€23,330/ QALY	€80,000/ QALY
Country – 3L								
United Kingdom	0.396	0.748	0.928	0.984	0.215	0.275	0.367	0.675
Spain	0.396	0.718	0.904	0.976	0.215	0.274	0.364	0.669
Japan	0.396	0.682	0.890	0.983	0.215	0.242	0.284	0.445
Zimbabwe	0.396	0.569	0.751	0.930	0.215	0.245	0.290	0.477
Germany	0.396	0.683	0.887	0.981	0.215	0.266	0.344	0.631
United States	0.396	0.599	0.786	0.932	0.215	0.261	0.331	0.603
Netherlands	0.396	0.711	0.906	0.982	0.215	0.267	0.348	0.635
South Korea	0.396	0.603	0.800	0.956	0.215	0.248	0.298	0.505
Denmark	0.396	0.674	0.873	0.972	0.215	0.258	0.324	0.576
France	0.396	0.752	0.919	0.974	0.215	0.271	0.356	0.635
Thailand	0.396	0.761	0.938	0.988	0.215	0.255	0.316	0.542
Canada	0.396	0.648	0.853	0.970	0.215	0.249	0.230	0.505
China	0.396	0.642	0.840	0.959	0.215	0.252	0.307	0.526
Italy	0.396	0.669	0.866	0.968	0.215	0.259	0.324	0.574
Singapore	0.396	0.770	0.918	0.966	0.215	0.287	0.396	0.703
Taiwan	0.396	0.810	0.952	0.987	0.215	0.274	0.362	0.641
Country – 5L								
United Kingdom	0.396	0.744	0.925	0.948	0.215	0.275	0.366	0.670
Spain	0.396	0.713	0.901	0.975	0.215	0.274	0.363	0.664
Japan	0.396	0.678	0.887	0.983	0.215	0.243	0.283	0.451
Zimbabwe	0.396	0.565	0.745	0.927	0.215	0.246	0.291	0.478
Germany	0.396	0.679	0.884	0.980	0.215	0.266	0.343	0.625
United States	0.396	0.595	0.780	0.929	0.215	0.261	0.330	0.597
Netherlands	0.396	0.707	0.903	0.981	0.215	0.268	0.347	0.629
South Korea	0.396	0.598	0.795	0.955	0.215	0.247	0.294	0.490
Denmark	0.396	0.669	0.870	0.971	0.215	0.259	0.325	0.577
France	0.396	0.752	0.974	0.975	0.215	0.257	0.319	0.541
Thailand	0.396	0.757	0.936	0.988	0.215	0.255	0.315	0.538
Canada	0.396	0.643	0.848	0.969	0.215	0.249	0.299	0.500
China	0.396	0.637	0.836	0.958	0.215	0.252	0.306	0.521
Italy	0.396	0.664	0.862	0.967	0.215	0.258	0.325	0.569
Singapore	0.396	0.766	0.916	0.965	0.215	0.288	0.394	0.700
Taiwan	0.396	0.806	0.951	0.986	0.215	0.273	0.360	0.680

QALY indicates quality-adjusted life year.

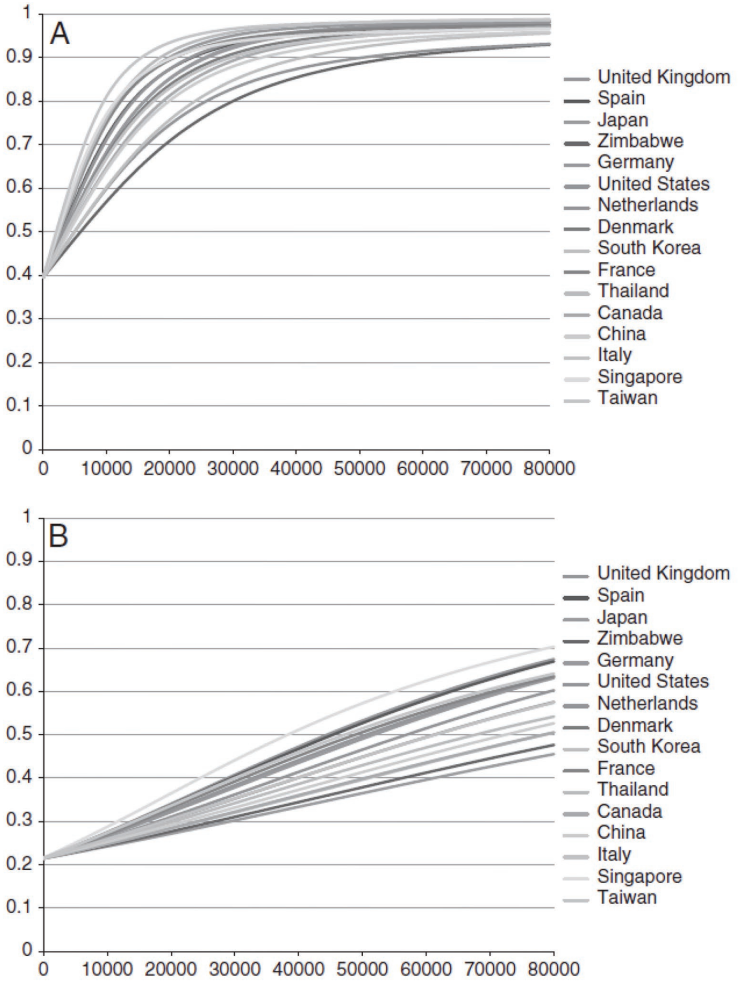


Figure 1 | A) Cost-effectiveness acceptability curves developed using the various country-specific value sets for the 3L- based on the low back pain study. B) Cost-effectiveness acceptability curves developed using the various country-specific value sets for the 3L-based on the depression study.

Discussion

This study explored the impact of using different country-specific EQ-5D value sets on cost-utility outcomes by employing 16 different EQ-5D country-specific value sets for both the 3L and 5L. Results showed that the use of different value sets may impact on cost-utility outcomes. To illustrate, for the 3L, ICERs ranged between €2044/QALY (Taiwan) and €5897/QALY (Zimbabwe) in the LBP study and between €38,287/QALY (Singapore) and €96,550/QALY (Japan) in the depression study. Moreover, at the lower bound of the Dutch QALY threshold, the probability of a treatment-based classification system for subacute and chronic LPB patients being cost-effective compared with usual care was found to range between 0.569 in Zimbabwe and 0.810 in Taiwan.

Previous studies on the impact of country-specific EQ-5D value sets were mostly based on the 3L.[13,18,53–55] In line with the present findings, these studies found Danish, German, Dutch, United States, and Japanese utility values derived from the 3L to be generally higher than those of the United Kingdom.[13,53,55] Also, Badia et al.[18] and Kiadaliri et al.[56] found Spanish and Swedish 3L utility values to be either higher or lower than those of the United Kingdom, depending of the severity of a health state. Another study of Lien et al.[54] found that United States, Danish, French, German, Japanese, and Dutch ICERs derived from the 3L differed extensively with that of Canada, with relative differences ranging from −17% (Germany) to +16% (United Kingdom). Similar results have recently been found for the 5L, when using both crosswalks and available 5L value sets.[19,57] To the best of our knowledge, the impact of country-specific EQ-5D value sets on the probability of an intervention being cost-effective compared with the control has never been explored, whereas this is one of the most important outcomes of a cost-utility analysis for decision-making purposes.

Strengths of this study include the fact that it was the first to systematically assess the impact of country-specific EQ-5D value sets on the probability of an intervention being cost-effective, its inclusion of a broad range of countries, and its use of empirical study data of patients suffering from both a physical and a mental health problem. Also, special efforts were made to ensure that the observed differences across countries could be ascribed to sociocultural differences instead of methodological ones. This was done by excluding value sets derived using elicitation methods other than the TTO, value sets that were derived from populations other than the general population, and value sets that contained interaction terms, quadratic variables, and/or additional health states, such as death or unconscious. By using the crosswalk approach, we were able to assess the impact of country-specific EQ-5D value sets for both the 3L and the 5L. Even though evidence suggests that utility values derived using the nonparametric crosswalk approach cannot be used interchangeably with those derived using available 5L value sets, both have previously been found to be associated with similar differences across countries.[19,57] This suggests that our use of crosswalks as a proxy of actual 5L utility values provides important preliminary evidence that using different country-specific EQ-5D value sets may impact on cost-utility outcomes. Further research using actual 5L utility values is needed to confirm this.

Several limitations are noteworthy as well. First, even though special efforts were made to ensure comparability across countries with regards to the methods they used to elicit population preferences, the impact of using different value sets could also depend on factors other than sociocultural differences, such as (a) the applied modeling technique and (b) the quality of the underlying data.[57,58] The way in which these factors impact on cost-utility outcomes should, therefore, be explored in a future study. Second, in both example studies, some cost and EQ-5D descriptive system data were missing. Missing data were handled using multiple imputation under the assumption that they were only related to observed data and not to unobserved data (ie, Missing At Random). As missing data are common in trial-based economic evaluations and multiple imputation is the recommended approach for dealing with missing data in such studies we do not expect the presence of missing data to have severely biased our results.[59,60] Also, it is important to mention that the present cost-utility analyses were only intended to illustrate the possible implications of the use of different country-specific value sets and not to serve as a bona fide cost-utility analysis of the interventions under study. For the latter, we refer to previous publications.[22,23]

Third, both example studies included data from 2 relatively small Dutch randomized controlled trials with specific patient populations and relatively few patients presenting more severe EQ-5D health states. Therefore, our findings should be viewed as an important indication that the use of different EQ-5D country-specific value sets might result in different cost-utility outcomes and further research is needed to investigate whether these results are generalizable to other studies, countries, and/or patient groups. Fourth, the nonparametric crosswalk of van Hout et al.[47] is based on self-reported health status of 3L and 5L versions of the EQ-5D in responders of 6 European countries, which include Denmark, the United Kingdom, Italy, the Netherlands, Poland, and Scotland. Therefore, it may not adequately capture differences in 5L utility values that might exist for other countries. To assess the possible impact of this issue, a sensitivity analysis was performed, in which the transition probabilities of van Hout et al.[47] were multiplied by 5L utility values, instead of 3L utility values, for countries that already have a 5L value set.[47] As the results of this sensitivity analysis were in line with those of the main analysis, we do not expect our reliance on the nonparametric crosswalk to have severely biased our results.

This study indicates that the use of different EQ-5D country-specific value sets has an impact on cost-utility outcomes, for both the 3L and 5L. This indicates that cost-utility outcomes may not be directly transferable across countries. Transferring economic evaluation results across countries is sometimes necessary, because many jurisdictions request information on the cost-effectiveness of interventions, while it is not possible to provide this information for every jurisdiction separately.[61] As country-specific value sets are thought to better reflect sociocultural differences,[13,61] this does not mean that more generic value sets, such as the European one,[62] are preferred. Instead, researchers are encouraged to conduct sensitivity analyses to assess the impact of using different country-specific value sets on their results and to develop methods for transferring cost-utility outcomes across countries. Health care decision-makers, on the other hand, are recommended to interpret cost-utility analysis results from other countries with caution.

Until now, various studies have been performed to assess the transferability of cost-utility outcomes across countries and/or to develop methods for improving their transferability.[61,63–65] These studies indicate that various factors influence the transferability of cost-utility outcomes across countries, including not only differences in value sets, but also differences in resource use patterns, unit costs, and baseline risks. Up until now, the majority of research focused on the cost side of the equation.[61,63–65] Recently, however, the "ISPOR Good Research Practices Task Force" emphasized the importance of the fact that utilities are not directly transferable across countries as well;[61] something which was confirmed by the present study. Oddershede and colleagues were the first to develop a method for converting Dutch and German 3L utility values into United Kingdom ones.[61,63–65] This set of analyses, however, only included a limited number of countries and did not account for the severity level of a health state. Therefore, more research into this area is warranted, particularly because the existence of successful methods for transferring cost-utility outcomes across countries could reduce the number of required studies and ensures that health care decisions can be made in a much more timely manner.[64]

In conclusion, this set of analyses indicates that the use of different EQ-5D country-specific value sets have an impact on cost-utility outcomes. Therefore, to account for the fact that health state preferences are affected by sociocultural differences, relevant country-specific value sets should be used in cost-utility analyses. Health care decision-makers are recommended to interpret cost-utility outcomes from other countries with caution and researchers are encouraged to develop methods for successfully transferring cost-utility outcomes from one country to another.

References

1. Karimi M, Brazier J. Health, health-related quality of life, and quality of life: what is the difference? Pharmacoeconomics. 2016;34:645–649.
2. Brazier J, Ratcliffe J, Salomon J, et al. Measuring and Valuing Health Benefits for Economic Evaluation, 2nd ed. Oxford, New York: Oxford University Press; 2016:372.
3. Finch AP, Brazier JE, Mukuria C. What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. Eur J Health Econ. 2017;19:1–14.
4. Brazier J, Ara R, Rowen D, et al. A review of generic preference-based measures for use in cost-effectiveness models. Pharmacoeconomics. 2017;35:21–31.
5. EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy. 1990;16:199–208.
6. Brazier JE, Harper R, Jones NM, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. BMJ. 1992;305:160–164.
7. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002;40:113–128.
8. Hawthorne DG, Je P, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of Health-Related Quality of Life. Qual Life Res. 1999;8:209–224.
9. Sintonen H. The 15-D measure of health related quality of life: reliability, validity and sensitivity of its health state descriptive system. Ann Med. 2001;33:328–336.
10. Torrance GW. Measurement of health state utilities for economic appraisal: a review. J Health Econ. 1986;5:1–30.
11. Stolk EA, Oppe M, Scalone L, et al. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. Value Health. 2010;13:1005–1113.
12. Krabbe PF, Devlin NJ, Stolk EA, et al. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Med Care. 2014;52:935–943.
13. Norman R, Cronin P, Viney R, et al. International comparisons in valuing EQ-5D health states: a review and analysis. Value Health. 2009;12:1194–1200.
14. Xie F, Gaebel K, Perampaladas K, et al. Comparing EQ-5D valuation studies: a systematic review and methodological reporting checklist. Value Health. 2013;16:A44–A45.
15. Brazier J, Roberts J, Tsuchiya A, et al. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ. 2004;13:873–84.
16. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20:1727–1736.
17. Bailey H, Kind P. Preliminary findings of an investigation into the relationship between national culture and EQ-5D value sets. Qual Life Res. 2010;19:1145–1154.
18. Badia X, Roset M, Herdman M, et al. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Making. 2001;21:7–16.
19. Mulhern B, Feng Y, Shah K, et al. Comparing the UK EQ-5D-3L and English EQ-5D-5L Value Sets. PharmacoEconomics. 2018;36:699–713.
20. Apeldoorn AT, Bosmans JE, Ostelo RW, et al. Cost-effectiveness of a classification-based system for sub-acute and chronic low back pain. Eur Spine J. 2012;21:1290–300.
21. Biesheuvel-Leliefeld KEM, Bosmans JE, Dijkstra-Kersten SMA, et al. A supported self-help for recurrent depression in primary care; an economic evaluation alongside a multi-center randomised controlled trial. PLoS One. 2018;13:e0208570.

22. Apeldoorn AT, Ostelo RW, van Helvoirt H, et al. The cost-effectiveness of a treatment-based classification system for low back pain: design of a randomised controlled trial and economic evaluation. BMC Musculoskelet Disord. 2010;11:58.

23. Biesheuvel-Leliefeld KE, Kersten SM, van der Horst HE, et al. Cost-effectiveness of nurse-led self-help for recurrent depression in the primary care setting: design of a pragmatic randomised controlled trial. BMC Psychiatry. 2012;12:59.

24. Bernert S, Fernández A, Haro JM, et al. Comparison of different valuation methods for population health status measured by the EQ-5D in three European countries. Value Health. 2009;12:750–758.

25. Dolan PDp. Modeling valuations for EuroQol health states : medical care. Med Care. 1997;35:1095–1108.

26. Devlin NJ, Shah KK, Feng Y, et al. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ. 2018;27:7–22.

27. Ramos-Goñi JM, Craig BM, Oppe M, et al. Handling data quality issues to estimate the Spanish EQ-5D-5L value set using a hybrid interval regression approach. Value Health. 2018;21:596–604.

28. Shiroiwa T, Ikeda S, Noto S, et al. Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. Value Health. 2016;19:648–654.

29. Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating an EQ-5D population value set: the case of Japan. Health Econ. 2002;11:341–353.

30. Jelsma J, Hansen K, de Weerdt W, et al. How do Zimbabweans value health states? Popul Health Metrics. 2003;1:11.

31. Ludwig K, Graf von der Schulenburg J-M, Greiner W. German value set for the EQ-5D-5L. PharmacoEconomics. 2018;36:663–674.

32. Claes C, Greiner W, Uber A, et al. An i\nterview-based comparison of the TTO and VAS values given to EuroQol states of health by the general German population. Centre for Health Economics and Health Systems Research, University of Hannover, Germany: EuroQol; 1999.

33. Shaw JW, Pickard AS, Yu S, et al. A median model for predicting United States population-based EQ-5D health state preferences. Value Health. 2010;13:278–288.

34. Versteegh MM, Vermeulen KM, Evers SMAA, et al. Dutch tariff for the five-level version of EQ-5D. Value Health. 2016;19:343–352.

35. Lamers LM. The transformation of utilities for health states worse than… : medical care. Med Care. 2007;45:238–244.

36. Kim S-H, Ahn J, Ock M, et al. The EQ-5D-5L valuation study in Korea. Qual Life Res. 2016;25:1845–1852.

37. Jo M-W, Yun S-C, Lee S-I. Estimating quality weights for EQ-5D health states with the time trade-off method in South Korea. Value Health. 2008;11:1186–1189.

38. Wittrup-Jensen KU, Lauridsen J, Gudex C, et al. Generation of a Danish TTO value set for EQ-5D health states. Scand J Public Health. 2009;37:459–466.

39. Chevalier J, de Pouvourville G. Valuing EQ-5D using time trade-off in France. Eur J Health Econ. 2013;14:57–66.

40. Pattanaphesaj J, Thavorncharoensap M, Ramos-Goñi JM, et al. The EQ-5D-5L valuation study in Thailand. Expert Rev Pharmacoecon Outcomes Res. 2018;18:551–558.

41. Tongsiri S, Cairns J. Estimating population-based values for EQ-5D health states in Thailand. Value Health. 2011;14:1142–1145.

42. Bansback N, Tsuchiya A, Brazier J, et al. Canadian valuation of EQ-5D health states: preliminary value set and considerations for future valuation studies. PLoS One. 2012;7:e31115.

43. Liu GG, Wu H, Li M, et al. Chinese time trade-off values for EQ-5D health states. Value Health. 2014;17:597–604.

44. Scalone L, Cortesi PA, Ciampichini R, et al. Italian population-based values of EQ-5D health states. Value Health. 2013;16:814–822.

45. Luo N, Wang P, Thumboo J, et al. Valuation of EQ-5D-3L health states in Singapore: modeling of time trade-off values for 80 empirically observed health states. Pharmacoeconomics. 2014;32:495–507.

46. Lee H-Y, Hung M-C, Hu F-C, et al. Estimating quality weights for EQ-5D (EuroQol-5 dimensions) health states with the time trade-off method in Taiwan. J Formos Med Assoc. 2013;112:699–706.

47. van Hout B, Janssen MF, Feng Y-S, et al. Interim Scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012;15:708–715.

48. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30:377–399.

49. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res. 2005;14:1523–1532.

50. Luo N, Johnson JA, Coons SJ. Using instrument-defined health state transitions to estimate minimally important differences for four preference-based health-related quality of life instruments. Med Care. 2010;48:365–371.

51. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. Health Econ. 2004;13:461–475.

52. Fenwick E, Claxton K, Sculpher M. Representing uncertainty: the role of cost-effectiveness acceptability curves. Health Econ. 2001;10:779–787.

53. Johnson JA, Luo N, Shaw JW, et al. Valuations of EQ-5D health states: are the United States and United Kingdom different? Med Care. 2005;43:221–228.

54. Lien K, Tam VC, Ko YJ, et al. Impact of country-specific EQ-5D-3L tariffs on the economic value of systemic therapies used in the treatment of metastatic pancreatic cancer. Curr Oncol. 2015;22:443.

55. Karlsson JA, Nilsson J-A, Neovius M, et al. National EQ-5D tariffs and quality-adjusted life-year estimation: comparison of UK, US and Danish utilities in south Swedish rheumatoid arthritis patients. Ann Rheum Dis. 2011;70:2163–2166.

56. Kiadaliri AA, Eliasson B, Gerdtham U-G. Does the choice of EQ-5D tariff matter? A comparison of the Swedish EQ-5D-3L index score with UK, US, Germany and Denmark among type 2 diabetes patients. Health Qual Life Outcomes. 2015;13:145.

57. Yang F, Devlin N, Luo N. Cost-utility analysis using EQ-5D-5L data: does how the utilities are derived matter? Value Health. 2019;22:45–49.

58. Olsen JA, Lamu AN, Cairns J. In search of a common currency: a comparison of seven EQ-5D-5L value sets. Health Econ. 2018;27:39–49.

59. Ramsey SD, Willke RJ, Glick H, et al. Cost-effectiveness analysis alongside clinical trials II—an ISPOR good research practices task force report. Value Health. 2015;18:161–172.

60. MacNeil Vroomen J, Eekhout I, Dijkgraaf MG, et al. Multiple imputation strategies for zero-inflated cost data in economic evaluations: which method works best? Eur J Health Econ. 2016;17:939–950.

61. Drummond M, Barbieri M, Cook J, et al. Transferability of economic evaluations across jurisdictions: ISPOR good research practices task force report. Value Health. 2009;12:409–418.

62. Greiner W, Weijnen T, Nieuwenhuizen M, et al. A single European currency for EQ-5D health states. Eur J Health Econ. 2003;4:222–231.

63. Oddershede L, Petersen KD. Adjustment of foreign EQ-5D-3L utilities can increase their transferability. Clinicoecon Outcomes Res. 2015;7:629–636.

64. Goeree R, He J, O'Reilly D, et al. Transferability of health technology assessments and economic evaluations: a systematic review of approaches for assessment and application. Clinicoecon Outcomes Res. 2011;3:89–104.

65. Welte R, Feenstra T, Jager H, et al. A decision chart for assessing and improving the transferability of economic evaluation results between countries. PharmacoEconomics. 2004;22:857–876.

Appendix 1 | Baseline characteristics of the example studies' participants

Baseline characteristics	**Low Back Pain study**		**Depression study**	
	Control group (n=82)	**Intervention group (n=74)**	**Control group (n=124)**	**Intervention group (n=124)**
Age (years) [mean, SD]	42 (10.9)	43 (11.7)	49 (11.5)	49 (12)
Gender [n, %]				
Female	49 (60%)	40 (54%)	84 (68%)	89 (72%)
Male	33 (40%)	34 (46%)	40 (32%)	35 (28%)
Level of education [n, %]				
Low	13 (16%)	14 (19%)	80 (65%)	71 (57%)
Middle	38 (46%)	23 (31%)	N.A.	N.A.
High	31 (38%)	37 (50%)	44 (35%)	53 (43%)
Perceived recovery (Yes) [n, %]	36 (44%)	45 (61%)	N.A.	N.A.
Marital status [n, %]				
Partner	63 (73%)	56 (76%)	80 (65%)	80 (65%)
No partner	25 (26%)	17 (23%)	44 (35%)	44 (35%)
Missing	1 (1%)	1 (1%)	0 (0%)	0 (0%)
Previous complaints (Yes) [n, %]	66 (81%)	68 (92%)	N.A.	N.A.
Treated per protocol (Yes) [n, %]	76 (93%)	66 (89%)	N.A.	N.A.
Type of complaints [n, %]				
Acute	19 (23%)	13 (18%)	N.A.	N.A.
Chronic	63 (77%)	61 (82)	N.A.	N.A.
Onset (years) [mean, SD]	N.A.	N.A.	27 (12.4)	28 (12)
Antidepressants [n, %]				
Yes	N.A.	N.A.	53 (53%)	50 (40%)
No	N.A.	N.A.	43 (35%)	48 (39%)
Missing	N.A.	N.A.	28 (12%)	26 (21%)
Utility [mean, SD]	0.77 (0.21)	0.83 (0.15)	0.78 (0.19)	0.77 (0.21)

Appendix 2 | Overview of the example studies' participants EQ-5D health states at baseline.

	Low Back Pain study (n=156)	**Depression study (n=248)**
Mobility [n (%)]		
No problems	111 (71%)	195 (79%)
Some problems	43 (28%)	40 (16%)
Extreme problems	1 (1%)	3 (1%)
Missing	*1 (1%)*	*10 (4%)*
Self-care [n (%)]		
No problems	124 (79%)	225 (91%)
Some problems	32 (21%)	13 (5%)
Extreme problems	0 (0%)	0 (0%)
Missing	*0 (0%)*	*10 (4%)*
Usual activities [n (%)]		
No problems	58 (37%)	132 (53%)
Some problems	94 (60%)	98 (40%)
Extreme problems	4 (3%)	8 (3%)
Missing	*0 (0%)*	*10 (4%)*
Pain/ discomfort [n (%)]		
No problems	25 (16%)	101 (41%)
Some problems	115 (74%)	126 (51%)
Extreme problems	16 (10%)	11 (4%)
Missing	*0 (0%)*	*10 (4%)*
Anxiety/ depression [n (%)]		
No problems	122 (78%)	128 (54%)
Some problems	32 (21%)	105 (42%)
Extreme problems	2 (1%)	5 (2%)
Missing	*0 (0%)*	*10 (4%)*

Appendix 3 | EQ-5D value sets for the different countries

	UK[1, 2]		Spain[3, 4]		Japan[5, 6]		Zimbabwe[7]		Germany[8, 9]		US[10]		Netherlands[11, 12]		South Korea[13, 14]		Denmark[15]		France[16]		Thailand[17, 18]		Canada[19]		China[20]		Italy[21]		Singapore[22]		Taiwan[23]	
	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L
Full health	1	1	1	1	1	1	0,9	N.A.	0,999	1	1	N.A.	1	1	1	1	1	N.A.	1	N.A.	1	1	1	N.A.	1	N.A.	1	N.A.	1	N.A.	1	N.A.
N1 constant	-0.081	N.A.	-0.024	N.A.	-0.152	-0.061	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	-0.071	-0.047	-0.050	-0.096	-0.114	N.A.	-0.081	N.A.	-0.202	N.A.	-0.111	N.A.	-0.041	N.A.	-0.043	N.A.	N.A.	N.A.	-0.185	N.A.
N3 constant	-0.269	N.A.	-0.291	N.A.	N.A.	N.A.	N.A.	N.A.	-0.323	N.A.	N.A.	N.A.	-0.234	N.A.	-0.050	N.A	N.A.	N.A.	-0.174	N.A.	-0.139	N.A.	N.A.	N.A.	-0.014	N.A.	N.A.	N.A.	-0.290	N.A.	-0.190	N.A.
N4 constant																-0.078																
Mobility																																
		-0.058		-0.084		-0.065		N.A.		-0.026		N.A.		-0.035		-0.046		N.A.		N.A.		-0.066		N.A.		N.A.		N.A.		N.A.		N.A.
	-0.069	-0.076	-0.106	-0.099	-0.075	-0.113	-0.056	N.A.	-0.099	-0.042	-0.04	N.A.	-0.036	-0.057	-0.096	-0.058	-0.053	N.A.	-0.115	N.A.	-0.121	-0.087	-0.046	N.A.	-0.097	N.A.	-0.076	N.A.	-0.168	N.A.	-0.123	N.A.
		-0.207		-0.25		-0.179		N.A.		-0.139		N.A.		-0.166		-0.133		N.A.		N.A.		-0.211		N.A.		N.A.		N.A.		N.A.		N.A.
	-0.314	-0.274	-0.430	-0.337	-0.418	-0.240	-0.204	N.A.	-0.327	-0.224	-0.49	N.A.	-0.161	-0.203	-0.418	-0.251	-0.411	N.A.	-0.372	N.A.	-0.432	-0.371	-0.322	N.A.	-0.249	N.A.	-0.518	N.A.	-0.304	N.A.	-0.272	N.A.
Self-care																																
		-0.05		-0.05		-0.038		N.A.		-0.050		N.A.		-0.038		-0.032		N.A.		N.A.		-0.058		N.A.		N.A.		N.A.		N.A.		N.A.
	-0.104	-0.08	-0.134	-0.053	-0.054	-0.070	-0.092	N.A.	-0.087	-0.056	-0.057	N.A.	-0.082	-0.061	-0.046	-0.05	-0.063	N.A.	-0.212	N.A.	-0.121	-0.071	-0.071	N.A.	-0.104	N.A.	-0.100	N.A.	-0.161	N.A.	-0.167	N.A.
		-0.164		-0.164		-0.118		N.A.		-0.169		N.A.		-0.168		-0.078		N.A.		N.A.		-0.192		N.A.		N.A.		N.A.		N.A.		N.A.
	-0.214	-0.203	-0.309	-0.196	-0.102	-0.161	-0.231	N.A.	-0.174	-0.260	-0.356	N.A.	-0.152	-0.168	-0.136	-0.122	-0.192	N.A.	-0.326	N.A.	-0.242	-0.250	-0.224	N.A.	-0.212	N.A.	-0.289	N.A.	-0.346	N.A.	-0.276	N.A.
Usual activities																																
		-0.05		-0.044		-0.057		N.A.		-0.036		N.A.		-0.039		-0.021		N.A.		N.A.		-0.058		N.A.		N.A.		N.A.		N.A.		N.A.
	-0.036	-0.063	-0.071	-0.049	-0.044	-0.092	-0.043	N.A.	N.A.	-0.049	-0.056	N.A.	-0.032	-0.087	-0.051	-0.051	-0.048	N.A.	-0.156	N.A.	-0.059	-0.071	-0.072	N.A.	-0.073	N.A.	-0.085	N.A.	-0.255	N.A.	-0.085	N.A.
		-0.162		-0.135		-0.155		N.A.		-0.129		N.A.		-0.192		-0.1		N.A.		N.A.		-0.154		N.A.		N.A.		N.A.		N.A.		N.A.
	-0.094	-0.184	-0.195	-0.153	-0.133	-0.173	-0.135	N.A.	N.A.	-0.209	-0.136	N.A.	-0.057	-0.192	-0.208	-0.175	-0.144	N.A.	-0.189	N.A.	-0.118	-0.248	-0.105	N.A.	-0.197	N.A.	-0.198	N.A.	-0.321	N.A.	-0.208	N.A.
Pain/ discomfort																																
		-0.063		-0.078		-0.041		N.A.		-0.057		N.A.		-0.066		-0.042		N.A.		N.A.		-0.056		N.A.		N.A.		N.A.		N.A.		N.A.
	-0.123	-0.084	-0.089	-0.101	-0.080	-0.068	-0.067	N.A.	-0.112	-0.109	-0.042	N.A.	-0.086	-0.092	-0.037	-0.053	-0.062	N.A.	-0.112	N.A.	-0.072	-0.067	-0.045	N.A.	-0.092	N.A.	-0.098	N.A.	-0.146	N.A.	-0.121	N.A.
		-0.276		-0.245		-0.124		N.A.		-0.404		N.A.		-0.36		-0.166		N.A.		N.A.		-0.207		N.A.		N.A.		N.A.		N.A.		N.A.
	-0.386	-0.335	-0.261	-0.382	-0.194	-0.193	-0.302	N.A.	-0.315	-0.612	-0.466	N.A.	-0.329	-0.415	-0.151	-0.207	-0.396	N.A.	-0.265	N.A.	-0.209	-0.256	-0.298	N.A.	-0.242	N.A.	-0.334	N.A.	-0.229	N.A.	-0.261	N.A.
Anxiety/ depression																																
		-0.078		-0.081		-0.078		N.A.		-0.030		N.A.		-0.07		-0.033		N.A.		N.A.		-0.058		N.A.		N.A.		N.A.		N.A.		N.A.
	-0.071	-0.104	-0.062	-0.128	-0.063	-0.111	-0.046	N.A.	N.A.	-0.082	-0.061	N.A.	-0.124	-0.145	-0.043	-0.046	-0.068	N.A.	-0.090	N.A.	-0.032	-0.096	-0.063	N.A.	-0.086	N.A.	-0.095	N.A.	-0.150	N.A.	-0.154	N.A.
		-0.285		-0.27		-0.173		N.A.		-0.244		N.A.		-0.356		-0.102		N.A.		N.A.		-0.233		N.A.		N.A.		N.A.		N.A.		N.A.
	-0.236	-0.289	-0.144	-0.348	-0.112	-0.197	-0.173	N.A.	-0.065	-0.356	-0.357	N.A.	-0.325	-0.421	-0.158	-0.137	-0.367	N.A.	-0.204	N.A.	-0.110	-0.295	-0.280	N.A.	-0.210	N.A.	-0.213	N.A.	-0.278	N.A.	-0.282	N.A.

CHAPTER 5

Can EQ-5D-3L utility values of low back pain patients be validly predicted by the Oswestry Disability Index for use in cost-effectiveness analyses?

Sylvia Pellekooren, Ângela Jornada Ben, Judith E. Bosmans, Raymond W. J. G. Ostelo, Maurits W. van Tulder, Esther T. Maas, Frank J.P.M. Huygen, Teddy Oosterhuis, Adri T. Apeldoorn, Miranda L. van Hooff, Johanna M. van Dongen

Quality of Life Research, 2022; 31: 2153 – 2165

Abstract

Purpose

To assess whether regression modelling can be used to predict EQ-5D-3L utility values from the Oswestry Disability Index (ODI) in low back pain (LBP) patients for use in cost effectiveness analysis.

Methods

EQ-5D-3L utility values of LBP patients were estimated using their ODI scores as independent variables using regression analyses, while adjusting for case-mix variables. Six different models were estimated: 1) Ordinary Least Squares (OLS) regression, with total ODI score, 2) OLS, with ODI item scores as continuous variables, 3) OLS, with ODI item scores as ordinal variables, 4) Tobit model, with total ODI score, 5) Tobit model, with ODI item scores as continuous variables, 6) Tobit model, with ODI item scores as ordinal variables. The models' performance was assessed using explained variance (R^2) and Root Mean Squared Error (RMSE). The potential impact of using predicted instead of observed EQ-5D-3L utility values on cost-effectiveness outcomes was evaluated in two empirical cost-effectiveness analysis.

Results

Complete individual patient data of 18,692 low back pain patients were analysed. All models had a more or less similar R^2 (range: 45–52%) and RMSE (range: 0.21–0.22). The two best performing models produced similar probabilities of cost-effectiveness for a range of willingness-to-pay (WTP) values compared to those based on the observed EQ-5D-3L values. For example, the difference in probabilities ranged from 2% to 5% at a WTP of 50,000 €/QALY gained.

Conclusion

Results suggest that the ODI can be validly used to predict low back pain patients' EQ-5D-3L utility values and QALYs for use in cost-effectiveness analyses.

Introduction

Low Back Pain (LBP) has an estimated incidence of 250 million people worldwide and is characterized by a high burden of disease.[1] Patients with LBP typically experience difficulties in different aspects of health-related quality of life, such as their daily functioning, social participation,[2,3] and working ability.[4,5] These difficulties may affect patients' health-related quality of life considerably,[3,6] and have a significant impact on healthcare and societal costs.[7,8] As limited (healthcare) resources are available, decision-makers are not only interested in the effectiveness of LBP treatments recommended in international guidelines, but also in their cost-effectiveness compared to alternative treatments.

Cost-effectiveness analysis provides insight into the relative cost-effectiveness of treatments by comparing their incremental costs to their incremental effects.[9] These effects are often expressed in Quality-Adjusted Life-Years (QALYs), which combine both the quality and quantity of life into a single outcome.[10] For estimating QALYs, health-related quality of life is typically measured using preference-based quality-of-life measures. Health states obtained from these measures can be converted into utility values, which represent the preferences of the general population of a country for given health states.[11] In many countries, it is recommended to estimate utility values using the EuroQol five-dimension questionnaire (EQ-5D) and national tariffs to account for the fact that health state preferences differ across countries.[12-14] Unfortunately, EQ-5D data are not always available in clinical trials,[15] as higher priority is sometimes given to condition-specific measures that assess more clinically relevant outcomes.[16]

When utility values are missing, QALYs cannot be calculated. However, information about the incremental cost per QALY gained is typically required by healthcare decision-makers, particularly at the national level.[12,13] In the absence of the EQ-5D or another generic preference-based quality-of-life measure, a condition-specific measure might be used to predict utility values.[17] In LBP, one of the most frequently used condition-specific measures is the Oswestry Disability Index (ODI).[18] The ODI measures limitations of a patient's performance,[19] and is recommended in the core outcome set for clinical trials in nonspecific LBP[20] and management of LBP.[21]

A previous study assessed the predictive ability of the ODI in estimating utility values from the EQ-5D-3L by using data from 14,544 patients with lumbar degenerative pathology treated in a tertiary spine centre.[22] Linear regression analysis was performed to predict the patients' EQ-5D utility values based on their ODI total or individual item scores and patients reported severity of back and leg pain. Based on a root mean squared error (RMSE) of 0.14, the authors concluded that it is not possible to estimate EQ-5D-3L utility values based on the ODI. However, given the bounded nature of EQ-5D data as well as the possible existence of other contextual factors that influence health-related quality of life in LBP, it is likely that the models' performance might be improved by using a Tobit model to account for possible ceiling effects. The model's performance might also be improved by including a wider variety of LBP patients treated in various settings, while adjusting for more case-mix variables. Moreover, the authors only based their conclusions on the models' RMSE without assessing the impact of using predicted utility scores in cost-effectiveness. Therefore, this study aimed to assess the feasibility of using different regression models to predict EQ-5D-3L utility values in LBP patients based on the ODI in cost-effectiveness analyses while adjusting for a broad range of case-mix characteristics.

Methods

Source of data

Individual patient data included in this study originated from four previously conducted prospective studies; i.e., the minimal interventional treatments (MINT) study, the rehabilitation after lumbar disc surgery (REALISE) study, the Nijmegen Decision Tool study, and a study evaluating a treatment-based classification system.[23-32] These studies were conducted among sub-acute and chronic LBP patients treated in primary care, secondary care, and/or tertiary care. For all patients, various sociodemographic variables were assessed at baseline, and both the ODI and EQ-5D-3L utility values were assessed at baseline and at one or more follow-up moments. In total, 21,500 patients were included in these studies. For developing the models, only baseline data were used in the present study, because the proportion of participants with missing data was low at baseline (i.e., <5%), thereby preventing the need for imputation of missing values. To assess the final models' performance in a trial-based cost effectiveness analysis setting, baseline as well as follow-up data were used of the MINT study,[23-25] and the treatment-based classification system study.[29,30]

The MINT study,[23-25] the REALISE study,[31,32] and the treatment-based classification system study[29,30] obtained ethical approval from the Medical Ethics Committee of the Erasmus Medical Centre Rotterdam or Medical Ethics Committee of the VU University Medical Centre in Amsterdam. For the Nijmegen Decision Tool study,[26-28] ethical approval was not required, because the *"Dutch Act on Medical Research involving Human Subjects"* does not apply to screening questionnaires that are part of routine practice. More detailed information on the design and study population of the different studies is provided in Appendix A.

Utility values

Utility values were based on the EQ-5D-3L, which is a generic preference-based measure that asks participants to describe their health state on five health dimensions (i.e., mobility, self-care, usual activities, pain/ discomfort, and anxiety/depression) using three severity levels (i.e., no problems, moderate problems, and severe problems).[33] The participants' EQ-5D-3L health states were converted into utility values using the Dutch tariff.[34] Utility values are presented on a continuous scale that is anchored at 1 (indicating full health) to 0 (indicating a state as bad as being dead). Negative values may also occur, which represent health states that are regarded as worse than a state that is as bad as being dead 10. Dutch EQ-5D-3L utility values can range between -0.33 and 1.

Oswestry Disability Index

The ODI measures the limitations of a patient's performance compared with that of a fit person, and consists of ten items assessing various aspects of daily living (e.g. lifting, walking, and travelling). Each item is scored on a six-point scale, ranging from 0 to 5. The overall ODI score was estimated by summing the values of all individual items, subsequently dividing this score by the total possible score, and multiplying this score by 100. The total score ranges from 0 to 100%, with higher scores indicate higher level of disability.[19,35] For this study, the "sex life" (item 8) was not included, as this item is frequently omitted in applied studies as well.[36-38] Including this item would have hampered the generalization of the results to a large number of LBP studies. The cross cultural adapted Dutch language version of the ODI version 2.1a was used in all studies included.[39]

Predictors

The following case-mix variables were included; age (years), gender (male/female), education level (low/moderate/high), living together with a partner (yes/no), type of LBP (sub-acute/chronic), setting (primary care/secondary care/ tertiary care), and back pain (Numeric Rating Scale (NRS: 0–10) Pain score: low 0–3, moderate 4–6, and severe 7–10).[40,41] These variables were included, because they were expected to increase the predictive value of the models[42-47] and to be measured in most applied studies, thereby increasing applicability of the models.

Statistical analysis

Baseline characteristics were described using frequencies and percentages for categorical variables and means and standard deviations for continuous variables. Prior to the development of the models, linearity and additivity assumptions (i.e., normally distributed residuals, homoscedasticity, influential cases and outliers) were assessed using diagnostic plots (i.e., scatterplot, density plot, and boxplots), and diagnostic tests (e.g., Grubbs test). Pearson's correlation coefficient was used to assess the strength of the linear relationship between the patients' EQ-5D-3L based utility values and ODI total scores. To assess the agreement between the EQ-5D-3L and the ODI the Intra Class Correlation (ICC) was calculated using a two-way random effects model.

Model development and variable selection

Models were developed using two regression techniques; i.e., Ordinary Least Squares (OLS) regression and Tobit regression (i.e. censored or truncated regression). OLS regression was included, because it is still one of the most frequently used linear modelling techniques. OLS regression is used to estimate the strength of the association between a continuous outcome variable and one or more independent variables.[48] OLS, however, does not take into account the bounded nature of utility values which can be accounted for in a Tobit regression.[49] This model can estimate linear relationships between variables, where the range of the dependent variable is constrained. This is done using a so-called latent variable that accounts for the fact that the true independent variable is – in our case – bounded at 1. Hereby, biased and inconsistent estimates, that may occur when using OLS regression, may be prevented.[50]

For both the OLS and Tobit model, three different regression models were developed; 1) including the overall ODI score as independent variable, 2) using all nine ODI items scores as independent variables and assuming them to be continuous, and 3) using all nine ODI items scores as independent variables and assuming them to be ordered. This resulted in six different models; 1) OLS, with the total ODI score, 2) OLS, with the ODI item scores as continuous variables, 3) OLS, with the ODI item scores as ordinal variables, 4) Tobit model, with the total ODI score, 5) Tobit model, with the ODI item scores as continuous variables, 6) Tobit model, with the ODI item scores as ordinal variables. To assess which variables increased the predictive value of the models, a bi-directional stepwise selection procedure,[51] using Akaike Information Criterion (i.e., the trade-off between the goodness of fit of the model and the simplicity of the model),[52] with a 5% significance level was used. Stepwise selection combines the elements of forward and backward selection by sequentially adding variables, based on the most contributing predictors, and omitting variables that no longer provide an improvement in the model fit after adding a new variable to the model. Final models only included case-mix variables that increased the predictive value.

Model performance and internal validation

The original dataset was split into a training sample (70%), and a validation sample (30%) using the 'create Data Partition' function in R. This function creates a balanced split of the data by performing a stratified random split of the data based on the mean of the dependent variable, which leads to a comparable mean EQ-5D-3L utility value in both the training and validation dataset. After developing the models in the training sample, their performance was assessed in the validation sample using the RMSE (i.e., the absolute fit of the model) and the adjusted R^2 (i.e., the relative fit of the model). The minimal important difference (MID) of the EQ-5D-3L was used to determine an acceptable RSME, which was set at a cut of point of 0.0353. A correlation of 0.5 or higher (i.e., a relatively moderate correlation as the R squared indicates that about half of the variance of the utility values is explained by the ODI) was considered sufficient for performing regression analysis. Recommended models were selected based on parsimony, which is the trade-off between between simplicity of the model (i.e., low AIC) and explanatory predictive power (i.e., high R^2). To assess agreement between the actual and estimated EQ-5D-3L based utility values a Bland Altman analysis was performed for all models.

Sensitivity Analyses

In addition to the main analysis, three sensitivity analyses (SA) were performed. In the first sensitivity analysis (SA1) the variable mental health status was added to the case-mix variables (SA1). SA1 was only performed on a sub-set of the data, as only one of the four datasets (i.e., the MINT study23-25) assessed mental health using the Four Dimensional Symptom Questionnaire (4DSQ),[53] and only part of the sample (n=4,123) completed this questionnaire. The 4DSQ assesses four different aspects of mental health (i.e., distress, depression, anxiety, and somatisation), all of which were included in the models as a separate variable. In SA2, the variable living with a partner was omitted. In SA3 the patients' EQ-5D-3L utility values were converted to EQ-5D-5L utility values using the reverse crosswalk (SA3).[55] Reversed cross walk values make it possible to link EQ-5D-3L responses to EQ-5D-5L value sets, and can be used when 5L values are wanted but only 3L data is available.[55,56] The 5-level EQ-5D version is an adapted version of the EQ-5D-3L, which is known to be more sensitive and has less ceiling effects, including through changing the number of levels of perceived problems per dimension from 3 to 5.[57]

Cost-effectiveness analysis

To assess the models' impact on cost-effectiveness outcomes, complete cases from two randomized controlled trials were used, i.e., empirical dataset 1 (n=68; Apeldoorn et al.[29-30]) and empirical dataset 2 (n=424; Maas et al.[23-25]). In both studies, QALYs were estimated based on both the actual EQ-5D-3L scores (i.e., actual QALY values) and based on the patients' ODI scores (i.e., predicted QALY values). Agreement between the actual and estimated EQ-5D-3L based utility values was assessed by performing a Bland Altman analysis for each of the empirical datasets.

Then, full trial-based cost effectiveness analyses were conducted for each of the six models as well as the patients' actual QALY values (i.e., QALYs based on the measured EQ-5D-3L scores). For each trial-based cost effectiveness analysis, mean differences in costs and QALYs between treatment groups were estimated using seemingly unrelated regression analyses. Incremental Cost-Effectiveness Ratios (ICERs) were calculated by dividing the difference in costs by the difference in effects. Uncertainty around cost and QALY differences was estimated using bootstrapping. The percentage of bootstrapped cost-effect pairs was reported per quadrant of the Cost-Effectiveness Plane (i.e., north-east, south-east, north-west, and south-west). Subsequently, Cost-Acceptability Curves (CEACs) were plotted. CEACs indicate an intervention's probability of cost-effectiveness compared to control for a range of willingness-to-pay (WTP) values (i.e., thresholds of 0, 30,000 euro and 50,000). These probabilities were assessed on their decision sensitivity (i.e., how sensitive is the conclusion of a cost effectiveness analysis is to using a particular statistical method).[58] Analyses were performed in R software, version 3.4.0.

Results

Participants

Out of the individual patient data that included 21,500 patients, 18,692 complete cases were included for analysis. These patients had sub-acute (n=3248) or chronic LBP (n=15,444). The mean age of the patients was 53.9 years (SD=14.7, range 18.1-91.9) and 61% of the sample was female. The patients' mean ODI score at baseline was 41.23 (SD=15.4, range 0-100) and their mean baseline EQ-5D-3L based utility value was 0.46 (SD=0.29, range -0.3290-1.00). More details on the patients' characteristics are shown in Table 1.

Variables included and model performance

The diagnostic plots showed a linear relationship between EQ-5D-3L based utility values and the ODI, and homogeneity of variance of the residuals. Even though the patients' baseline EQ-5D-3L based utility values followed a bimodal distribution, the corresponding residuals were normally distributed. Hence, the normality of residuals assumption of linear regression was met. No outliers or influential cases were identified. Pearson's correlation coefficient between the patients' baseline EQ-5D-3L utility values and ODI total score was 0.63. The ICC showed an agreement of 0.23 between individual ODI items and EQ-5D-3L items.

Table 1 | Baseline characteristics of included patients

Characteristic	n=18,692
Age (mean (SD), range)	53.9 (14.7), 18.1-91.9
Gender; female (n, %)	11,345 (60.7)
Education (n, %)	
Low (no education, primary level education, lower vocational and lower secondary education)	5,398 (28.9)
Moderate (higher secondary education or undergraduate)	9,078 (48.6)
High (tertiary, university level, postgraduate)	4,216 (22.6)
Living with a partner (n, %)	14,085 (75.4)
Type of LBP (n, %)	
Subacute (<3 months)	3,248 (17.4)
Chronic (>3 months)	15,444 (82.6)
Post-surgery (n, %)	1,587 (8.5)
Setting (n, %)	
Primary care (i.e., physiotherapy clinics)	150 (0.8)
Secondary care (i.e., pain clinics)	4,123 (22.1)
Tertiary care (i.e., hospital)	14,419 (77.1)
NRS Pain (mean (SD))	6.99 (1.9)
Utility score (mean (SD), range)	0.467 (0.299), -0.3290-1.00
ODI score[a] (mean (SD), range)	41.23 (15.4), 0-100
ODI 1 mean (SD)/ median (IQR)	2.66 (0.93) / 3 (2-4)
ODI 2 mean (SD)/ median (IQR)	1.11 (1.04) / 1 (0-2)
ODI 3 mean (SD)/ median (IQR)	2.78 (1.32) / 3 (2-4)
ODI 4 mean (SD)/ median (IQR)	1.44 (1.22) / 1 (0-2)
ODI 5 mean (SD)/ median (IQR)	2.11 (1.09) / 2 (1-3)
ODI 6 mean (SD)/ median (IQR)	2.85 (1.29) /3 (2-4)
ODI 7 mean (SD)/ median (IQR)	1.49 (1.09) / 1 (0-2)
ODI 9 mean (SD)/ median (IQR)	2.14 (1.20) / 2 (1-3)
ODI 10 mean (SD)/ median (IQR)	1.98 (1.32) / 2 (1-3)

[a] excluding item 8 sex life

LBP= Low Back Pain; NRS= Numeric Rating Scale (range 0-10); Utility (range -0.33 to 1); ODI= Oswestry Disability Scale (range 0-100); ODI individual item (range 0-5) ; SD= Standard Deviation IQR= Inter Quartile Range

An overview of the independent variables that were included in the final models, as well as their respective regression coefficients, can be found in Appendix B. The case-mix variables age, gender, education, partner and NRS were included in all models, whereas type of LBP was not included in any of the models. The variable setting was included in all models except for model 1 (i.e., OLS with ODI total scores). In the models using Tobit regression, 74 of the 13,087 observations in the training set were right censored.

The performance of the different models was more or less the same, with explained variances ranging from 45% to 51% and RMSEs ranging from 0.21 to 0.22. Based on parsimony of the models, model 2 and 5 seem most appropriate to use. More details on the performance of the different models are shown in Table 2.

The mean difference between estimated and actual utility values for model 2 was -0.068 (95% CI -0.495, 0.359), and for model 5 -0.086 (95%CI -0.512, 0.341). Bland Altman plots of models 2 and 5 are shown in Figure 1. The plots for other all models are presented in Appendix C.

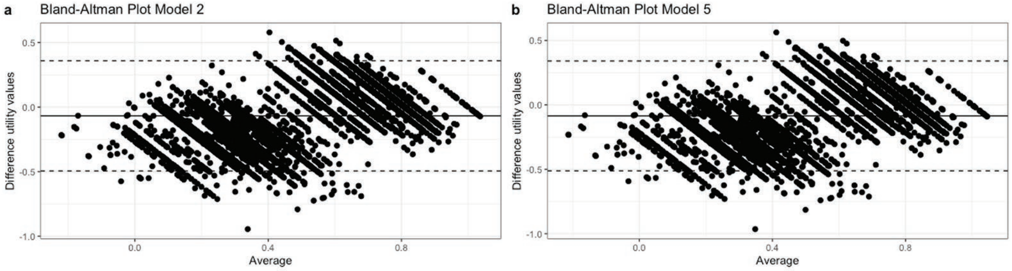


Figure 1 | Bland Altman plots models 2 and 5 – validation set

X-axis: Average measurement of the estimated and actual utility values, Y-axis: Difference in measurements between the two instruments

Solid line: Average difference in measurements between the estimated and actual utility values, Dashed lines: 95% confidence interval limits for the average difference

Table 2 | Performance measures in the training and validation sets

	Performance in the training set (n=13,087)			**Performance in validation set (n=5,605)**		
	R^2	RMSE	AIC	R^2	RMSE	AIC
Model 1: OLS with ODI total scores	0.45	0.22	-2326.48	0.46	0.22	-1083.26
Model 2: OLS with ODI individual item total scores continuous	0.50	0.21	-3423.24	0.50	0.21	-1513.73
Model 3: OLS with ODI individual item total scores ordered	0.51	0.21	-3769.51	0.52	0.21	-1638.09
Model 4: Tobit with ODI total scores	0.45	0.22	-2061.91	0.46	0.22	-951.61
Model 5: Tobit with ODI individual item total scores continuous	0.50	0.21	-3164.37	0.50	0.21	-1385.32
Model 6 Tobit with individual item total scores ordered	0.51	0.21	-3474.88	0.52	0.21	-1494.06

OLS: Ordinary Least Squares Regression, ODI: Oswestry Disability Index, R^2: proportion of variance for the dependent variable, RMSE: Root Mean Squared Error, AIC: Akaike Information Criteria.

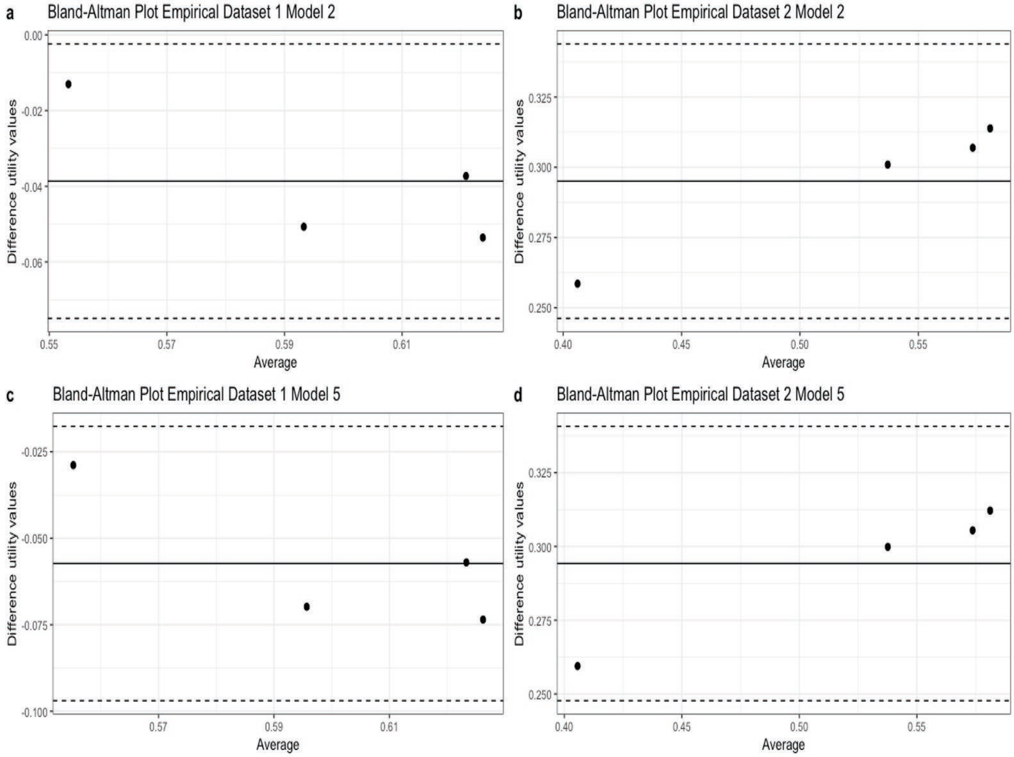


Figure 2 | Bland Altman plots models 2 and 5 – empirical datasets

X-axis: Average measurement of the estimated and actual utility values, Y-axis: Difference in measurements between the two instruments.

Solid line: Average difference in measurements between the estimated and actual utility values, Dashed lines: 95% confidence interval limits for the average difference.

Sensitivity analysis

Adding mental health variable(s) to the models resulted in an increase of the explained variance of 2–4%, whereas the RMSE remained similar. Omission of the variable 'living with a partner' (SA2) did not change the models' performance. Using the patients' reversed cross-walked EQ-5D-5L utility values (SA3) improved the models' explained variance by 3–4%, and the RMSE reduced with 0.06-0.07. More details on the results of the sensitivity analyses are provided in Appendix D.

Results cost-effectiveness analysis

The mean difference between estimated and actual utility values for empirical dataset 1 model 2 was -0.039 (95% CI -0.075, -0.002), and for model 5 -0.057 (95% CI -0.097, -0.018). The mean difference between estimated and actual utility values for empirical dataset 2 model 2 was 0.295 (95% CI 0.246, 0.344), and for model 5 the mean difference was 0.294 (95% CI 0.248, 0.341). Bland Altman plots of models 2 and 5 for both empirical datasets are shown in Figure 2. The plots for other all models are presented in Appendix E.

In both empirical datasets, the difference between the predicted and actual differences in QALYs was small for the two most parsimonious models (i.e. models 2 and 5:Δ≤0.004) and the distributions of cost-effect pairs across the four quadrants of the cost-effectiveness plane were comparable. The cost-effectiveness acceptability curves based on both predicted and actual QALY values were also similar. The predicted probability of an intervention being cost-effective at a willingness to pay of 50,000 was slightly higher in both models than the actual probabilities (i.e., 2–5% in model 2, and 3–5% in model 5). More details on the cost-effectiveness outcomes for all models in both empirical studies are shown in Table 3 and Figure 3.

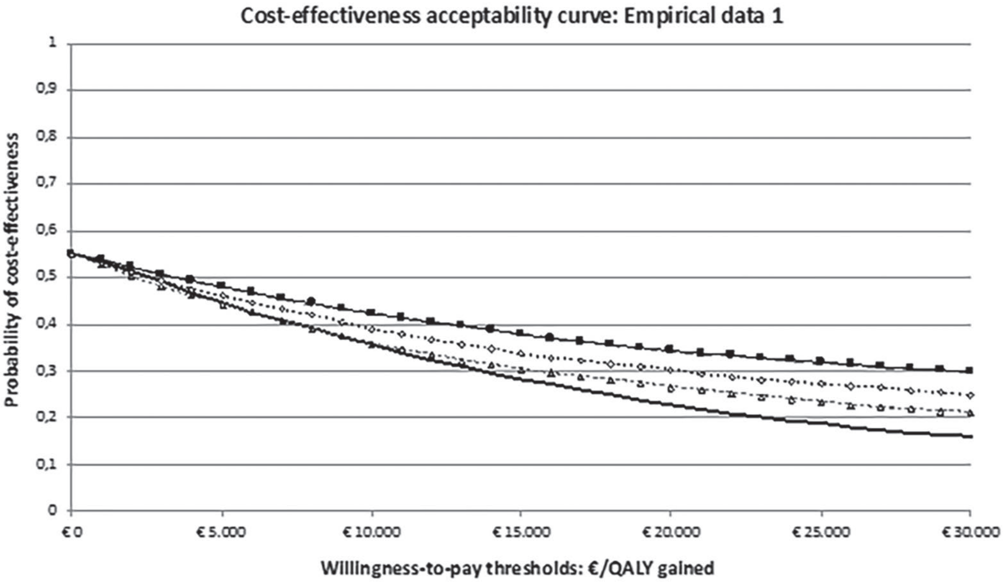


Figure 3 | Cost-effectiveness acceptability curves – empirical dataset 1

M1= Model 1; M2= Model 2; M3= Model 3; M4= model 4; M5= Model 5; M6= model 6

Table 3 | Cost-effectiveness outcomes for an intervention in comparison with usual care by predictive models

Predictive models	**ΔE** (95% CI)	**ΔC** (95% CI)	**ICER**	**Cost-effectiveness plane**				**Cost-effectiveness acceptability curve**			
				NE	**SE**	**SW**	**NW**	**P_{CE}(0)**	**P_{CE}(10,000)**	**P_{CE}(30,000)**	**P_{CE}(50,000)**
Empirical dataset 1 28, 29 N= 86											
Actual values	-0.041 (-0.091; 0.009)	-110 (-1761; 1283)	2697	2%	4%	51%	42%	0.55	0.36	0.16	0.11
Model 1	-0.035 (-0.094; 0.021)	-110 (-1761; 1283)	3091	1%	10%	45%	44%	0.55	0.39	0.25	0.20
Model 2	-0.043 (-0.106; 0.015)	-110 (-1761; 1283)	2559	1%	7%	48%	44%	0.55	0.36	0.21	0.16
Model 3	-0.027 (-0.081; 0.018)	-110 (-1761; 1283)	4068	1%	13%	43%	43%	0.55	0.42	0.30	0.24
Model 4	-0.036 (-0.095; 0.021)	-110 (-1761; 1283)	3058	1%	10%	45%	44%	0.55	0.39	0.25	0.20
Model 5	-0.044 (-0.107; 0.015)	-110 (-1761; 1283)	2514	1%	7%	48%	44%	0.55	0.36	0.21	0.16
Model 6	-0.027 (-0.080; 0.021)	-110 (-1761; 1283)	4084	2%	13%	42%	43%	0.55	0.42	0.30	0.25
Empirical dataset 2 22-24 N = 424											
Actual values	-0.004 (-0.034; 0.027)	1576 (596; 2575)	-371566	38%	0%	0%	62%	0.001	0.002	0.017	0.048
Model 1	-0.007 (-0.037; 0.023)	1576 (596; 2575)	-226441	32%	0%	0%	68%	0.001	0.002	0.014	0.037
Model 2	0.0002 (-0.030; 0.029)	1576 (596; 2575)	6670132	51%	0%	0%	49%	0.001	0.003	0.025	0.070
Model 3	-0.001 (-0.026; 0.024)	1576 (596; 2575)	-2099247	48%	0%	0%	52%	0.001	0.002	0.015	0.028
Model 4	-0.007 (-0.037; 0.024)	1576 (596; 2575)	-224080	32%	0%	0%	67%	0.001	0.002	0.014	0.038
Model 5	0.0003 (-0.030; 0.030)	1576 (596; 2575)	5105447	51%	0%	0%	49%	0.001	0.003	0.025	0.073
Model 6	-0.001 (-0.027; 0.026)	1576 (596; 2575)	-2417793	48%	0%	0%	51%	0.001	0.002	0.018	0.053

Recommended models are presented as bold text. N= number of observations in the analysis; ΔC= difference in costs; 95% CI= 95% confidence interval; ΔE= difference in effects; ICER= Incremental Cost-Effectiveness Ratio; NE= northeast; SE= southeast; SW= southwest; NW= northwest; P_{CE}(0)= probability that the intervention is cost-effective as compared to usual care with a threshold of 0; P_{CE}( )= probability that the intervention is cost-effective as compared to usual care with willingness-to-pay thresholds of 0, 10,000, 30,000, and 50,000 Euros.

Discussion

Main findings

There were no large differences in the models' performance between OLS and Tobit regression, nor between using the patients' total ODI scores and ODI individual item scores. The explained variance of the developed models ranged from 45% to 51%, and the RMSE ranged from 0.21 to 0.22. Models 2 and 5 are recommended based on the best fit and parsimony. The models' relatively low absolute fit (RMSE) indicates that they are not suitable for estimating utility values for individual patients. Nonetheless, they can be used to predict differences in LBP patients' EQ-5D-3L utility values and QALY's, as the systematic bias in mean scores does not affect the differences between the groups. Cost-effectiveness outcomes of models 2 and 5 based on predicted and actual values were similar. These findings enable researchers to perform a cost-effectiveness analysis with QALYs as the outcome measure, even if EQ-5D-3L data are missing.

Comparison with literature

Our findings regarding the performance measures are more or less in line with the previous study by Carreon et al.,[20] who aimed to predict individual LBP patients' EQ-5D-3L utility values based on their ODI scores. Their model performed slightly better in terms of its explained variance (i.e., R^2 was 61%) and its absolute fit (i.e., RMSE is 0.149), which is probably the result of a more homogenous study population, and therefore may indicate overfitting of their model. Based on the RMSE, Carreon et al.[20] concluded that individual patients' EQ-5D-3L utility values could not validly be predicted from their ODI scores. Although we agree with this conclusion, we would like to stress that a low RMSE does not necessarily mean that the models cannot be used in the context of a cost-effectiveness analysis. This is true when the bias surrounding the predicted utility values does not translate into relevant differences in incremental QALYs and the probability of the intervention being cost-effective compared to the control group (i.e., decision-based validity).[57] This may be explained by the fact that the bias is likely to be similar in the intervention and control groups, thereby not affecting incremental QALYs and CEACs.[59]

Strengths and limitations

To develop the models, a large sample of LBP patients from various settings (i.e., primary, secondary, and tertiary care) and with various complaint durations (i.e., subacute and chronic LBP) was used, which increases both the reliability and generalisability of the models. Moreover, next to OLS models, Tobit models were used to account for the constrained range of utility values.[49,50] Although the added value of the Tobit model in this LBP population turned out to be rather limited, this might be different for LBP populations with milder symptoms, in which a larger share of patients is expected to report full health (i.e., a utility value of 1).

Our study also had some limitations. First, part of the sample was derived from two RCTs. Although RCT data may have limited generalisability, we chose to add these RCTs to our sample to create a more diverse sample and provide a better representation of the LPB population. Second, during the analysis, balanced data splitting was used to create the training and validation set.

Although this balanced split provides better distribution of data then a random split, it might have been more appropriate to use K-fold cross validation.[60] Unfortunately, running the Tobit model using k-fold cross validation was not feasible as the R package for the Tobit model was not compatible with the K-fold package. In a post-hoc analysis we developed and validated the OLS models with k-fold cross validation and this produced similar results as our main analysis (data not shown). We also expect this to be the case for the Tobit models. Third, EQ-5D-3L utilities were used instead of EQ-5D-5L utilities. This is a limitation because EQ-5D-5L is known to be more sensitive and therefore recommended in pharmacoeconomic guidelines. Nonetheless, some countries still use the EQ-5D-3L. Therefore, we preferred to use the current relatively large dataset with EQ-5D-3L utility values of nearly 20,000 patients for developing and validating the models, instead of using a relatively small dataset with EQ-5D-5L. As the performance measures in the sensitivity analysis using the EQ-5D-5L reversed crosswalk were comparable with those of the EQ-5D-3L version, we expect that EQ-5D-5L values can also be validly estimated using ODI scores. Fourth, the models were based on Dutch utility values. Previous research[14] has shown that there are differences in utilities, QALYs, ICERs, and CEACs between countries due to the use of different value sets per country. Therefore, we added the regression coefficients of models 2 and 5 for different countries in Appendix F. These regression coefficients are based on the available value sets (tariffs) for different countries and can be used to calculate utility values and QALYs. Fifth, some data that were used to assess the performance of the developed models in a trial-based cost-effectiveness analysis setting were also part of the training set. However, as this was only a small percentage of the total training set (3.1%), we do not expect it to have influenced the validity of our finding that the difference between the estimated and true QALYs is small. Last, for assessing the performance of the developed models in a trial-based cost-effectiveness analysis setting, we only used data of two clinical trials, both of which found the intervention far from being cost-effective. That is, the probability of the interventions being cost-effective was low regardless of the willingness to pay threshold. In datasets where the interventions' cost-effectiveness is less conclusive, even small differences in the probability of an intervention being cost-effective might impact the overall conclusion of a study. Further research in the form of a simulation study, using simulated data to examine the generalisability beyond the datasets, is needed to assess the performance of the developed models in a wide range of trial-based cost-effectiveness analysis settings.

Implications for research and practice

Our findings suggest that predictive modelling can be used to estimate utility values from disease-specific measures, such as the ODI amongst LBP patients when assessing incremental costs per QALY gained (as part of a cost-effectiveness analysis) or differences in utilities between groups. This is helpful for assessing cost-effectiveness in trials that did not directly measure utilities. Given the relatively large RMSE (i.e., low absolute fit of the models) and the relatively low r-square value (i.e., low relative fit) it is strongly discouraged to use the developed models to estimate the utility values of individual patients. Further research is needed to validate the models in order to 1) assess whether these models yield comparable results in other empirical datasets on LBP interventions, especially in analysis of interventions that are expected not to be more conclusive in their cost-effectiveness, and 2) to improve their generalisability among different LBP patients by external validation in another sample. This study focussed on assessing the validity of predictive regression modelling in estimating EQ-5D-3L utility values from the ODI and the impact of these estimated utility values on cost-effectiveness analysis. Results show that this is feasible for estimating QALYs and ICERs, but not for estimating individual utility scores. Further research is needed to explore whether adjusted regression techniques, such as response mapping techniques like non-parametric and multinomial logistic regression,[16,54,55] result in better predictive accuracy in estimating individual utility values of preference-based measures, such as the EQ-5D. This is important because studies suggest these mapping methods might be better at preventing regression to the mean.[61] Additional research might not only result in more accurate estimated utility values but would also provide insight into the relative performance of different methods to estimate these values.

In the meantime, researchers can use the developed models in their cost-effectiveness analysis when utility values are lacking. Of them, the OLS model (i.e., model 2) is recommended in samples in which only a small number of patients has a utility value of 1 at baseline or follow-up measurement, whereas the Tobit model (i.e., model 5) is recommended in samples in which a substantial part of the sample has a utility score at baseline or at follow-up measurement. Although it seems possible to estimate utility values from disease-specific measures it is important to stress that it is still preferred to use preference-based quality of life measurements when setting up new studies.

Conclusions

Results of this study suggest that the ODI can be used to predict LBP patients' EQ-5D-3L utility values when the aim is to perform a cost-effectiveness analysis for QALYs, if utility values are missing, in order to compare the difference between groups of patients. The models are not suitable for estimating utility values for individual patients. Further research is needed to validate the models in order to assess whether these models yield comparable results in other empirical datasets on LBP interventions, to improve generalisability of the estimated models, and to compare the performance of predictive modelling compared to a mapping approach for estimating utility values. In the meantime, researchers can use the developed models in their cost-effectiveness analysis when utility values are lacking.

References

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet(London, England), 392*(10159), 1789–1858. https:// doi. org/ 10. 1016/ S0140- 6736(18) 32279-7
2. MacNeela, P., Doyle, C., O'Gorman, D., Ruane, N., & McGuire, B. E. (2015). Experiences of chronic low back pain: A meta-ethnography of qualitative research. *Health psychology review, 9*(1), 63–82. https:// doi. org/ 10. 1080/ 17437 199. 2013. 840951
3. Froud, R., Patterson, S., Eldridge, S., Seale, C., Pincus, T., Rajendran, D., Fossum, C., & Underwood, M. (2014). A systematic review and meta-synthesis of the impact of low back pain on people'slives. *BMC Musculoskeletal Disorders, 15*, 50. https:// doi.org/ 10. 1186/ 1471- 2474- 15- 50
4. Ihlebaek, C., Hansson, T. H., Laerum, E., Brage, S., Eriksen, H. R., Holm, S. H., Svendsrod, R., & Indahl, A. (2006). Prevalence of low back pain and sickness absence: A "borderline" study in Norway and Sweden. *Scandinavian Journal of Public Health, 34*(5), 555–558. https:// doi. org/ 10. 1080/ 14034 94060 05520 51
5. Steenstra, I. A., Munhall, C., Irvin, E., Oranye, N., Passmore, S., Van Eerd, D., Mahood, Q., & Hogg-Johnson, S. (2017). Systematic Review of Prognostic Factors for Return to Work in Workers with Sub Acute and Chronic Low Back Pain. *Journal of Occupational Rehabilitation, 27*(3), 369–381. https:// doi. org/ 10. 1007/s10926- 016- 9666-x
6. Hush, J. M., Refshauge, K., Sullivan, G., De Souza, L., Maher, C. G., & McAuley, J. H. (2009). Recovery: What does this mean to patients with low back pain? *Arthritis and Rheumatism, 61*(1), 124–131. https:// doi. org/ 10. 1002/ art. 24162
7. Maniadakis, N., & Gray, A. (2000). The economic burden of back pain in the UK. *Pain, 84*(1), 95–103. https:// doi. org/ 10. 1016/ S0304- 3959(99) 00187-6
8. Lambeek, L. C., van Tulder, M. W., Swinkels, I. C., Koppes, L. L., Anema, J. R., & van Mechelen, W. (2011). The trend in total cost of back pain in The Netherlands in the period 2002 to 2007. *Spine, 36*(13), 1050–1058. https:// doi. org/ 10. 1097/ BRS. 0b013 e3181 e70488
9. Drummond, M. F., Sculpher, M. J., Torrance, G. J., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes*. Oxford University Press.
10. Brazier, J., Ratcliffe, J., Saloman, J., & Tsuchiya, A. (2016). *Measuring and valuing health benefits for economic evaluation*. Oxford University Press.
11. Froberg, D. G., & Kane, R. L. (1989). Methodology for measuring health-state preferences—I: Measurement strategies. *Journal of Clinical Epidemiology, 42*(4), 345–354. https:// doi. org/ 10. 1016/0895- 4356(89) 90039-5
12. Hakkaart-van Roijen L., van der Linden N., Bouwmans C. A. M., Kanters T. A., & Tan S. S. (2015) *Kostenhandleiding: Methodologie van kostenonderzoek en referentieprijzen voor economische evaluaties in de gezondheidszorg*. Zorginstituut Nederland. Retrieved August 30, 2021, from https:// www. zorgi nstit uutne derland. nl/ over- ons/ publi caties/ publi catie/ 2016/ 02/ 29/ richt lijn- voorhet-uitvo eren-van- econo mische- evalu aties- in- de- gezon dheid szorg
13. National Institute for Health and Care Excellence. (2013). *Guide to the methods of technology appraisal.* Retrieved August 30, 2021, from https:// www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l
14. van Dongen, J. M., Jornada Ben, A., Finch, A. P., Rossenaar, M., Biesheuvel-Leliefeld, K., Apeldoorn, A. T., Ostelo, R., van Tulder, M. W., van Marwijk, H., & Bosmans, J. E. (2021). Assessing the impact of EQ-5D country-specific value sets on cost-utility outcomes. *Medical care, 59*(1), 82–90. https:// doi. org/ 10. 1097/MLR. 00000 00000 001417

15. Gianola, S., Frigerio, P., Agostini, M., Bolotta, R., Castellini, G., Corbetta, D., Gasparini, M., Gozzer, P., Guariento, E., Li, L. C., Pecoraro, V., Sirtori, V., Turolla, A., Andreano, A., & Moja, L. (2016). Completeness of outcomes description reported in low back pain rehabilitation interventions: A Survey of 185 randomized trials. *Physiotherapy Canada, 68*(3), 267–274. https://doi. org/ 10. 3138/ ptc. 2015- 30. PMID: 27909 376; PMCID: PMC5125456

16. Fayers, P. M., & Hays, R. D. (2014). Should linking replace regression when mapping from profile-based measures to preference-based measures? *Value in Health, 17*(2), 261–265. https://doi. org/ 10. 1016/j. jval. 2013. 12. 002

17. Longworth, L., & Rowen, D. (2013). Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. *Value in Health, 16*(1), 202–210. https:// doi. org/ 10. 1016/j. jval.2012. 10. 010

18. Chapman, J. R., Norvell, D. C., Hermsmeyer, J. T., Bransford, R. J., DeVine, J., McGirt, M. J., & Lee, M. J. (2011). Evaluating common outcomes for measuring treatment success for chronic low back pain. *Spine, 36*(21 Suppl), S54–S68. https:// doi. org/ 10.1097/ BRS. 0b013 e3182 2ef74d

19. Fairbank, J. C., Couper, J., Davies, J. B., & O'Brien, J. P. (1980). The Oswestry low back pain disability questionnaire. *Physiotherapy,66*(8), 271–273.

20. Chiarotto, A., Boers, M., Deyo, R. A., Buchbinder, R., Corbin, T. P., Costa, L., Foster, N. E., Grotle, M., Koes, B. W., Kovacs, F. M., Lin, C. C., Maher, C. G., Pearson, A. M., Peul, W. C., Schoene, M. L., Turk, D. C., van Tulder, M. W., Terwee, C. B., & Ostelo, R. W. (2018). Core outcome measurement instruments for clinicaltrials in nonspecific low back pain. *Pain, 159*(3), 481–495. https://doi. org/ 10. 1097/j. pain. 00000 00000 001117

21. Clement, R. C., Welander, A., Stowell, C., Cha, T. D., Chen, J. L., Davies, M., Fairbank, J. C., Foley, K. T., Gehrchen, M., Hagg, O., Jacobs, W. C., Kahler, R., Khan, S. N., Lieberman, I. H., Morisson, B., Ohnmeiss, D. D., Peul, W. C., Shonnard, N. H., Smuck, M. W., et al. (2015). A proposed set of metrics for standardized outcome reporting in the management of low back pain. *Acta Orthopaedica, 86*(5), 523–533. https://doi.org/10. 3109/17453674. 2015.10366 96

22. Carreon, L. Y., Bratcher, K. R., Das, N., Nienhuis, J. B., & Glassman, S. D. (2014). Estimating EQ-5D values from the Oswestry Disability Index and numeric rating scales for back and leg pain. *Spine, 39*(8), 678–682. https:// doi. org/ 10. 1097/ BRS. 00000 00000000220

23. Mutubuki, E. N., Luitjens, M. A., Maas, E. T., Huygen, F., Ostelo, R., van Tulder, M. W., & van Dongen, J. M. (2020). Predictive factors of high societal costs among chronic low back pain patients. *European Journal of Pain (London, England), 24*(2), 325–337. https:// doi. org/ 10. 1002/ ejp. 1488

24. Maas, E. T., Juch, J. N., Groeneweg, J. G., Ostelo, R. W., Koes, B. W., Verhagen, A. P., van Raamt, M., Wille, F., Huygen, F. J., & van Tulder, M. W. (2012). Cost-effectiveness of minimal interventional procedures for chronic mechanical low back pain: Design of four randomised controlled trials with an economic evaluation. *BMC Musculoskeletal Disorders, 13*, 260. https:// doi. org/ 10. 1186/1471- 2474- 13- 260

25. Juch, J., Maas, E. T., Ostelo, R., Groeneweg, J. G., Kallewaard, J.W., Koes, B. W., Verhagen, A. P., van Dongen, J. M., Huygen, F., & van Tulder, M. W. (2017). Effect of radiofrequency denervation on pain intensity among patients with chronic low back pain: The mint randomized clinical trials. *JAMA, 318*(1), 68–81. https:// doi.org/ 10. 1001/ jama. 2017. 7918

26. van Hooff, M. L., van Loon, J., van Limbeek, J., & de Kleuver, M. (2014). The Nijmegen decision tool for chronic low back pain. Development of a clinical decision tool for secondary or tertiary spine care specialists. *PLoS ONE*. https:// doi. org/ 10. 1371/ journal. pone. 01042 26

27. van Dongen, J. M., van Hooff, M. L., Spruit, M., de Kleuver, M., & Ostelo, R. (2017). Which patient-reported factors predict referral to spinal surgery? A cohort study among 4987 chronic low back pain patients. *European Spine Journal, 26*(11), 2782–2788. https:// doi. org/ 10. 1007/ s00586- 017- 5201-9

28. 28. van Hooff, M. L., van Dongen, J. M., Coupe, V. M., Spruit, M., Ostelo, R., & de Kleuver, M. (2018). Can patient-reported profiles avoid unnecessary referral to a spine surgeon? An observational study to further develop the Nijmegen Decision Tool for Chronic Low Back Pain. *PLoS ONE, 13*(9), e0203518. https:// doi. org/ 10.1371/ journ al. pone. 02035 18

29. Apeldoorn, A. T., Ostelo, R. W., van Helvoirt, H., Fritz, J. M., de Vet, H. C., & van Tulder, M. W. (2010). The cost-effectiveness of a treatment-based classification system for low back pain: Design of a randomised controlled trial and economic evaluation. *BMC Musculoskeletal Disorders, 11*, 58. https:// doi. org/ 10.1186/ 1471- 2474- 11- 58

30. Apeldoorn, A. T., Ostelo, R. W., van Helvoirt, H., Fritz, J. M., Knol, D. L., van Tulder, M. W., & de Vet, H. C. (2012). A randomized controlled trial on the effectiveness of a classification based system for subacute and chronic low back pain. *Spine, 37*(16), 1347–1356. https:// doi. org/ 10. 1097/ BRS. 0b013 e31824d9f2

31. Oosterhuis, T., van Tulder, M., Peul, W., Bosmans, J., Vleggeert-Lankamp, C., Smakman, L., Arts, M., & Ostelo, R. (2013). Effectiveness and cost-effectiveness of rehabilitation after lumbar disc surgery (REALISE): Design of a randomised controlled trial. *BMC Musculoskeletal Disorders, 14*, 124. https:// doi. org/ 10. 1186/ 1471- 2474- 14- 124

32. Oosterhuis, T., Ostelo, R. W., van Dongen, J. M., Peul, W. C., de Boer, M. R., Bosmans, J. E., Vleggeert-Lankamp, C. L., Arts, M. P., & van Tulder, M. W. (2017). Early rehabilitation after lumbar disc surgery is not effective or cost-effective compared to no referral: A randomised trial and economic evaluation. *Journal of Physiotherapy, 63*(3), 144–153. https:// doi. org/ 10.1016/j. jphys. 2017. 05. 016

33. EuroQol Group. (1990). EuroQol–a new facility for the measurement of health-related quality of life. *Health Policy (Amsterdam, Netherlands), 16*(3), 199–208. https:// doi. org/ 10. 1016/0168- 8510(90) 90421-9

34. Lamers, L. M., Stalmeier, P. F., McDonnell, J., Krabbe, P. F., & van Busschbach, J. J. (2005). Kwaliteit van leven meten in economische evaluaties: Het Nederlands EQ-5D-tarief [Measuring the quality of life in economic evaluations: The Dutch EQ-5D tariff]. *Nederlands Tijdschrift Voor Geneeskunde, 149*(28), 1574–1578.

35. Fairbank, J. C., & Pynsent, P. B. (2000). The Oswestry Disability Index. *Spine, 25*(22), 2940–2952. https:// doi. org/ 10. 1097/ 00007 632- 20001 1150- 00017

36. Hudson-Cook, N., Tomes-Nicholson, K., & Breen, A. A. (1989). Revised Oswestry disability questionnaire. In M. Roland & J. R. Jenner (Eds.), *Back pain: New approaches to rehabilitation and education* (pp. 187–204). Manchester University Press.

37. Yeomans, S. G., & Liebenson, C. (1997). Applying outcomes management to clinical practice. *Journal of the Neuromusculoskeletal System, 5*(1), 1–14.

38. Shearer H. M. (2007). Rehabilitation of the Spine—A practitioner's manual, 2nd Ed. *The Journal of the Canadian Chiropractic Association, 51*(1), 62.

39. van Hooff, M. L., Spruit, M., Fairbank, J. C., van Limbeek, J., & Jacobs, W. C. (2015). The Oswestry Disability Index (version 2.1a): validation of a Dutch language version. *Spine, 40*(2), E83–E90. https:// doi. org/ 10. 1097/ BRS. 00000 00000 000683

40. Downie, W. W., Leatham, P. A., Rhind, V. M., Wright, V., Branco, J. A., & Anderson, J. A. (1978). Studies with pain rating scales. *Annals of the Rheumatic Diseases, 37*(4), 378–381. https:// doi.org/ 10. 1136/ ard. 37.4. 378

41. Boonstra, A. M., Stewart, R. E., Koke, A. J., Oosterwijk, R. F., Swaan, J. L., Schreurs, K. M., & Schiphorst Preuper, H. R. (2016). Cut-off points for mild, moderate, and severe pain on the numeric rating scale for pain in patients with chronic musculoskeletal pain: Variability and influence of sex and catastrophizing. *Frontiers in Psychology, 7*, 1466. https:// doi. org/ 10. 3389/ fpsyg. 2016. 01466

42. Husky, M. M., Ferdous Farin, F., Compagnone, P., Fermanian, C., & Kovess-Masfety, V. (2018). Chronic back pain and its association with quality of life in a large French population survey. *Health and Quality of Life Outcomes, 16*(1), 195. https:// doi. org/10. 1186/ s12955- 018- 1018-4

43. Horng, Y. S., Hwang, Y. H., Wu, H. C., Liang, H. W., Mhe, Y. J., Twu, F. C., & Wang, J. D. (2005). Predicting health-related quality of life in patients with low back pain. *Spine, 30*(5), 551–555. https:// doi. org/ 10. 1097/ 01. brs. 00001 54623. 20778. f0

44. Kovacs, F. M., Abraira, V., Zamora, J., Fernandez, C., & Spanish Back Pain Research Network. (2005). The transition from acute to subacute and chronic low back pain: A study based on determinants of quality of life and prediction of chronic disability. *Spine, 30*(15), 1786–1792. https:// doi. org/ 10. 1097/ 01. brs. 00001 72159.47152. dc

45. Lame, I. E., Peters, M. L., Vlaeyen, J. W., Kleef, M., & Patijn, J. (2005). Quality of life in chronic pain is more associated with beliefs about pain, than with pain intensity. *European Journal of Pain (London, England), 9*(1), 15–24. https:// doi. org/ 10. 1016/j.ejpain. 2004. 02. 006

46. Bentsen, S. B., Wahl, A. K., Strand, L. I., & Hanestad, B. R. (2007). Relationships between demographic, clinical and pain variables and health-related quality of life in patients with chronic low back pain treated with instrumented fusion. *Scandinavian Journal of Caring Sciences, 21*(1), 134–143. https:// doi. org/ 10.1111/j. 1471- 6712. 2007. 00440.x

47. Coste, J., Lefrancois, G., Guillemin, F., Pouchot, J., & French Study Group for Quality of Life in Rheumatology. (2004). Prognosis and quality of life in patients with acute low back pain: Insights from a comprehensive inception cohort study. *Arthritis and Rheumatism, 51*(2), 168–176. https:// doi. org/ 10. 1002/ art. 20235

48. Hutcheson, G. D. (1999). *The multivariate social scientist*. SAGE Publications, Ltd.

49. Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrics, 26*(1), 24–36.

50. Austin, P. C., Escobar, M., & Kopec, J. A. (2000). The use of the Tobit model for analyzing measures of health status. *Quality of Life Research, 9*(8), 901–910. https:// doi. org/ 10. 1023/a: 1008938326 604

51. Smith, G. (2018). Step away from stepwise. *Journal of Big Data*. https:// doi. org/ 10. 1186/ s40537- 018- 0143-6

52. Chowdhury, M., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*. https:// doi. org/ 10. 1136/fmch- 2019- 000262

53. Soer, R., Reneman, M. F., Speijer, B. L., Coppes, M. H., & Vroomen, P. C. (2012). Clinimetric properties of the EuroQol-5D in patients with chronic low back pain. *The Spine Journal, 12*(11), 1035–1039. https://d oi.o rg/1 0.1 016/j.s pinee.2 012.1 0.0 30

54. Terluin, B., van Marwijk, H. W., Ader, H. J., de Vet, H. C., Penninx, B. W., Hermens, M. L., van Boeijen, C. A., van Balkom, A. J., van der Klink, J. J., & Stalman, W. A. (2006). The Four-Dimensional Symptom Questionnaire (4DSQ): A validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. *BMC Psychiatry, 6*, 34. https:// doi. org/ 10. 1186/ 1471- 244X-6- 34

55. van Hout, B. A., & Shaw, J. W. (2021). Mapping EQ-5D-3L to EQ-5D-5L. *Value in Health, 24*(9), 1285–1293. https:// doi. org/10. 1016/j. jval. 2021. 03. 009

56. Euroqol. *Cross-walk and reverse cross-walk*. Retrieved August 30, 2021, from https:// euroq ol. org/ suppo rt/ tools/ analy sis- tools/cross- walk- rever se- cross- walk/

57. Euroqol. *EQ-5D-5L version*. Retrieved August 30, 2021, from https:// euroq ol. org/ eq- 5d- instr uments/ eq- 5d- 5l- about/

58. Griffin, S. (2010). *Dealing with uncertainty in the economic evaluationof health care technologies*. University of York.

59. Ben, Â., Finch, A. P., van Dongen, J. M., de Wit, M., van Dijk, S., Snoek, F. J., Adriaanse, M. C., van Tulder, M. W., & Bosmans, J. E. (2020). Comparing the EQ-5D-5L crosswalks and value sets for England, the Netherlands and Spain: Exploring their impact on cost-utility results. *Health Economics, 29*(5), 640–651. https://doi. org/ 10. 1002/ hec. 4008

60. Blum, A., Kalai, A., & Langford, J. (1999). Beating the hold-out: Bounds for k-fold and progressive cross-validation. In Association for Computing Machinery, *Proceedings of the twelfth annual conference on Computational learning theory (COLT '99)*. (pp. 203–208).

61. Thompson, N. R., Lapin, B. R., & Katzan, I. L. (2017). Mapping PROMIS global health items to EuroQol (EQ-5D) utility scores using linear and equipercentile equating. *PharmacoEconomics, 35*(11), 1167–1176. https:// doi. org/ 10. 1007/ s40273- 017- 0541-1

Appendix A | Description Studies Included

MINT Study

The MINT study 23-25 assessed the effectiveness of radiofrequency denervation added to a standardized exercise program for patients with chronic low back pain. This study included patients with chronic LBP, receiving conservative treatment in a multidisciplinary pain clinic. This study was conducted at 16 multidisciplinary pain clinics in the Netherlands, and had both a randomized and observational track. The randomized track consisted of three sub trails, namely the facet joint trial, the sacroiliac joint trial, and the combination trial (facet joint, sacroiliac joint, or the intervertebral disk). Patients were consecutively screened, and were eligible when meeting the following criteria; a. pain considered to be related to the facet joint, sacroiliac joint, or a combination of the facet joint, sacroiliac joint, or intervertebral disk, aged 18 to 70 years, and no improvement in symptoms after conservative treatment. A total of 681 patients were included in the three randomized trails. Patients who were not willing to participate, or did not meet the inclusion criteria, were approached for the observational track of this study. In total 5168 patients were included in the observational track. Exclusion criteria for all trials were pregnancy, severe psychological problems, involvement in work-related conflicts or claims; body mass index higher than 35; or anticoagulant drug therapy or coagulopathy. Data was collection through surveys. For more details we refer to the original publications.

Nijmegen Decision Tool Study

In the Nijmegen Decision Tool Study (NDT study) 26-27 47 indicators for a successful treatment outcome were assessed among chronic low back pain patients (CLBP), in order to compile a decision-support screening tool (NDT-CLBP) 28. Patients were recruited at a Dutch orthopaedic hospital specialized in spine care, prior to their first consultation at the orthopaedic outpatient department. All consecutive low back pain patients were asked to complete the web-based questionnaire, which is part of routine practice. In total 14,859 patients with chronic LBP were included in this dataset. Patients were eligible when meeting the following criteria; experienced low back pain complaints for more than three months (i.e., CLBP) due to degenerative lumbar spine disorders (excluding trauma and tumor), had access to the internet, and were able to read and write Dutch. For more details we refer to the original publications.

Study of Apeldoorn et al.

The study of Apeldoorn et al. 29,30 assessed the cost-effectiveness of a modified version of Delitto's classification-based treatment approach compared with usual physical therapy care in patients with sub-acute and chronic LBP. This study included 156 patients with subacute and chronic LBP treated in a primary care setting. Patients were recruited by during their first contact with a physical therapist working in the region of Amsterdam. Patients were eligible when meeting the following criteria; LBP as the primary complaint (with or without associated leg pain), age between 18 and 65 years, current episode longer than 6 weeks, and able to read and write Dutch. Exclusion criteria were known- or suspected-specific LBP, severe radiculopathy, serious co-morbidity and psychopathology. Data was collection through surveys. For more details we refer to the original publications.

REALISE Study

The REALISE study 31-32 concerned the assessment of effectiveness, and cost effectiveness of referral for early rehabilitation after lumbar disc surgery. This multicentre, randomised, controlled trial included 169 LPB patients with a herniated lumbar disc postoperatively treated in a primary care facility. Patients were referred to the research team by neurosurgeons, and checked on eligibility by research nurses. Patients were eligible when meeting the following criteria; a herniated lumbar disc confirmed by magnetic resonance imaging (MRI) and signs of nerve root compression corresponding to the level of disc herniation, aged between 18 and 70 years, and were able to fill out questionnaires in Dutch themselves. Exclusion criteria were cauda equina syndrome, neurogenic claudication, co-morbidities of the lumbar spine, spinal surgery in the prior 12 months, contraindications to exercise therapy, pregnancy, or previous lumbar disc surgery at the same level and on the same side. Data was collection through surveys. For more details we refer to the original publications.

Table A1 | Baseline characteristics included studies (complete cases EQ-5D and ODI)

	Apeldoorn Study 29, 30 n=156	MINT Study 23-25 n=6,316	Nijmegen study 26-28 n=14,859	REALISE Study 31-32 n=169
Age (years), mean (SD)	42.5 (11.2)	56.2 (13.5)	53.5 (15.1)	47.3 (11.8)
Sex, n [%]				
Female	89 (57.1)	3,576 (67.1)	8,695 (58.5)	98 (58.0)
Male	67 (42.9)	1,757 (32.9)	6,164 (41.5)	71 (42.0)
Education level, n [%]				
Low (no education, primary level education, lower vocational and lower secondary education)	27 (17.3)	1,892 (29.9)	3,922 (26.4)	37 (21.9)
Moderate (higher secondary education or undergraduate)	61 (39.1)	2,406 (38.1)	6,967 (47.8)	97 (57.4)
High (tertiary, university level, postgraduate)	68 (43.6)	823 (13.03)	3,403 (22.9)	35 (20.7)
Living together with a partner, n [%]				
Yes	119 (76.3)	4,663 (73.8)	11,118 (74.8)	125 (74.0)
No	37 (23.7)	1,593 (25.2)	3,741 (25.2)	44 (26.0)
Type of low back pain, n [%]				
Subacute (<3 months)	32 (20.5)	3,601 (57.0)	423 (2.8)	0
Chronic (>3 months)	124 (79.5)	1,682 (26.6)	14,436 (97.2)	169 (100.0)
Post-surgery, n [%]				
Yes	0	0	0	169 (100.0)
No	156 (100.0)	6,316 (100.0)	14,859 (100.0)	0
Setting, n [%]				
Primary care (i.e., physiotherapy clinics)	156 (100.0)	0	0	169 (100.0)
Secondary care (i.e., pain clinics)	0	6,316 (100.0)	0	0
Tertiary care (i.e., hospital)	0	0	14,859 (100.0)	
NRS Pain, mean (SD)	6.1 (1.8)	7.3 (1.6)	6.9 (2.0)	6.3 (2.6)
Utility, mean (SD)	0.7 (0.2)	0.5 (0.3)	0.5 (0.3)	0.4 (0.3)
ODI, mean (SD)	20.6 (13.0)	39.6 (14.6)	42.0 (15.4)	31.1 (14.3)

SD= Standard Error, NRS= Numeric Rating Scale range 0-10, Utility range: -0.33 – 1, ODI: Oswestery Disability Index range: 0-100

Appendix B | Regression coefficients models 1–6

Model 1. Ordinary Least Squares Regression with ODI total scores

Utility = 0.833 - 0.011*ODI total score + 0.002*age + 0.012*female + 0.015 *education middle + 0.021 *education high - 0.014 *no partner + 0.015* NRS moderate - 0.115 *NRS severe

	Regression Coefficient (SE)	95% CI	
		2.5%	97.5%
Intercept	0.833 (0.016)	0.807	0.857
ODI total score	-0.011 (0.000)	-0.011	-0.010
Age	0.002 (0.000)	0.001	0.002
Gender; female	0.012 (0.004)	0.003	0.019
Education; middle	0.015 (0.004)	0.006	0.024
Education; high	0.021 (0.006)	0.010	0.032
Partner; no partner	-0.014 (0.005)	-0.023	-0.005
NRS; moderate	0.015 (0.008)	-0.001	0.031
NRS; severe	-0.115 (0.009)	-0.131	-0.098

Model 2. Ordinary Least Squares Regression with ODI individual items scores (continuous)

Utility = 0.936 - 0.095*ODI1 - 0.044*ODI2 - 0.005*ODI3 - 0.019*ODI4 - 0.004*ODI5 - 0.008*ODI6 - 0.014*ODI7 - 0.033*ODI9 - 0.019*ODI10 + 0.002*age + 0.008*female + 0.019*education middle + 0.026*education high - 0.014*no partner - 0.066* secondary care - 0.051*tertiary care + 0.034*NRS moderate - 0.062*NRS severe

	Regression Coefficient (SE)	95% CI	
		2.5%	97.5%
Intercept	0.936 (0.024)	0.889	0.984
ODI1	-0.095 (0.003)	-0.099	-0.089
ODI2	-0.044 (0.002)	-0.049	-0.039
ODI3	-0.005 (0.002)	-0.009	-0.002
ODI4	-0.019 (0.002)	-0.023	-0.015
ODI5	-0.004 (0.002)	-0.008	-0.000
ODI6	-0.008 (0.002)	-0.012	-0.005
ODI7	-0.014 (0.002)	-0.018	-0.010
ODI9	-0.033 (0.002)	-0.039	-0.029
ODI10	-0.019 (0.002)	-0.023	-0.015
Age	0.002 (0.000)	0.001	0.002
Gender; female	0.008 (0.004)	0.000	0.015
Education; middle	0.019 (0.005)	0.009	0.027
Education; high	0.026 (0.005)	0.016	0.037
Partner; no partner	-0.014 (0.004)	-0.023	-0.006
Setting; secondary care	-0.066 (0.022)	-0.108	-0.024
Setting; tertiary care	-0.051 (0.021)	-0.093	-0.009
NRS; moderate	0.034 (0.008)	0.019	0.049
NRS; severe	-0.062 (0.008)	-0.079	-0.045

Model 3. Ordinary Least Squares Regression with ODI individual items scores (ordered)

Utility = 0.794 + 0.020*ODI1;1 - 0.004*ODI1;2 -0.138*ODI1;3 - 0.246*ODI1;4 - 0.247*ODI1;5 - 0.053 *ODI2;1 + 0.006*ODI2;2 - 0.106*ODI2;3 - 0.190*ODI2;4 - 0.146*ODI2;5 + 0.001*ODI3;1 - 0.001*ODI3;2 - 0.006*ODI3;3 - 0.012*ODI3;4 - 0.039*ODI3;5 - 0.017*ODI4;1 - 0.033*ODI4;2 -0.048*ODI4;3 - 0.069*ODI4;4 - 0.131*ODI4;5 + 0.004*ODI5;1 + 0.006*ODI5;2 - 0.009*ODI5;3 - 0.016*ODI5;4 - 0.026*ODI5;5 - 0.001*ODI6;1 - 0.005*ODI6;2 - 0.011*ODI6;3 - 0.024*ODI6;4 - 0.043*ODI6;5 - 0.003*ODI7;1 - 0.023*ODI7;2 - 0.036 *ODI7;3 - 0.049*ODI7;4 - 0.051*ODI7;5 - 0.024*ODI9;1 - 0.034*ODI9;2 - 0.093*ODI9;3 - 0.154*ODI9;4 - 0.153*ODI9;5 - 0.020*ODI10;1 - 0.042*ODI10;2 - 0.060*ODI10;3 - 0.079*ODI10;4 - 0.066*ODI10;5 + 0.002*age + 0.007*female + 0.017*education middle + 0.026*education high - 0.013*no partner - 0.088*secondary care - 0.078*tertiary care + 0.002*NRS moderate - 0.080*NRS severe

	Regression Coefficient (SE)	95% CI	
		2.5%	97.5%
Intercept	0.794 (0.028)	0.738	0.849
ODI1;1	0.020 (0.018)	-0.015	0.055
ODI1;2	-0.004 (0.017)	-0.037	0.029
ODI1;3	-0.138 (0.017)	-0.172	-0.104
ODI1;4	-0.246 (0.018)	-0.281	-0.211
ODI1;5	-0.247 (0.022)	-0.291	-0.204
ODI2;1	-0.053 (0.005)	-0.063	-0.043
ODI2;2	0.006 (0.006)	-0.109	-0.087
ODI2;3	-0.106 (0.008)	-0.122	-0.089
ODI2;4	-0.190 (0.015)	-0.221	-0.160
ODI2;5	-0.146 (0.040)	-0.224	-0.070
ODI3;1	0.001 (0.010)	-0.019	0.020
ODI3;2	-0.001 (0.010)	-0.021	0.020
ODI3;3	-0.006 (0.010)	-0.025	0.013
ODI3;4	-0.012 (0.010)	-0.032	0.008
ODI3;5	-0.039 (0.013)	-0.064	-0.013
ODI4;1	-0.017 (0.005)	-0.026	-0.007
ODI4;2	-0.033 (0.006)	-0.045	-0.021
ODI4;3	-0.048 (0.007)	-0.062	-0.033
ODI4;4	-0.070 (0.010)	-0.088	-0.050
ODI4;5	-0.131 (0.028)	-0.186	-0.077
ODI5;1	0.004 (0.008)	-0.012	0.020
ODI5;2	0.006 (0.008)	-0.009	0.021
ODI5;3	-0.009 (0.009)	-0.025	0.008
ODI5;4	-0.016 (0.010)	-0.036	0.005
ODI5;5	-0.026 (0.018)	-0.061	0.008
ODI6;1	-0.001 (0.011)	-0.022	0.021
ODI6;2	-0.005(0.011)	-0.027	0.016
ODI6;3	-0.011(0.011)	-0.032	0.011
ODI6;4	-0.024 (0.011)	-0.045	-0.003
ODI6;5	-0.043(0.013)	-0.070	-0.017
ODI7;1	-0.003(0.006)	-0.015	0.009

	Regression Coefficient (SE)	95% CI	
		2.5%	97.5%
ODI7;3	-0.036(0.008)	-0.108	-0.078
ODI7;4	-0.049(0.013)	-0.074	-0.025
ODI7;5	-0.051(0.016)	-0.082	-0.020
ODI9;1	-0.024(0.007)	-0.039	-0.009
ODI9;2	-0.034(0.007)	-0.048	-0.019
ODI9;3	-0.093(0.008)	-0.108	-0.078
ODI9;4	-0.154(0.011)	-0.175	-0.132
ODI9;5	-0.153(0.011)	-0.187	-0.119
ODI10;1	-0.020(0.008)	-0.037	-0.004
ODI10;2	-0.042(0.009)	-0.060	-0.024
ODI10;3	-0.060(0.010)	-0.079	-0.040
ODI10;4	-0.079 (0.011)	-0.101	-0.057
ODI10;5	-0.066(0.012)	-0.090	-0.041
Age	0.002(0.000)	0.001	0.002
Gender; female	0.007(0.004)	-0.001	0.014
Education; middle	0.017(0.004)	0.008	0.026
Education; high	0.026(0.005)	0.015	0.036
Partner; No partner	-0.013(0.004)	-0.021	-0.004
Setting; secondary care	-0.088 (0.022)	-0.131	-0.046
Setting; tertiary care	-0.078 (0.022)	-0.120	-0.035
NRS; moderate	0.002 (0.008)	-0.014	0.018
NRS; severe	-0.080 (0.009)	-0.096	-0.063

Model 4. Tobit with ODI total scores

Utility = 0.897 - 0.011*ODI total score + 0.002*age + 0.011*female + 0.015*education middle + 0.021*education high - 0.014*no partner - 0.058*secondary care - 0.058*tertiary care + 0.010*NRS moderate - 0.119*NRS severe

	Regression Coefficient (SE)	95% CI	
		2.5%	97.5%
Intercept	0.897 (0.025)	0.848	0.947
ODI total score	-0.011 (0.000)	-0.011	-0.011
Age	0.002 (0.000)	0.002	0.002
Gender; female	0.011 (0.004)	0.004	0.020
Education; middle	0.015 (0.004)	0.006	0.024
Education; high	0.021 (0.006)	0.010	0.032
Partner; no partner	-0.014 (0.005)	-0.023	-0.005
Setting; secondary care	-0.058 (0.023)	-0.104	-0.013
Setting; tertiary care	-0.058 (0.023)	-0.103	-0.013
NRS; moderate	0.010 (0.008)	-0.006	0.026
NRS; severe	-0.119 (0.009)	-0.136	-0.102

Model 5. Tobit with ODI individual items scores (continuous)

Utility = 0.961 - 0.096*ODI1 - 0.044*ODI2 - 0.005*ODI3 - 0.019*ODI4 - 0.005*ODI5 - 0.009*ODI6 - 0.014*ODI7 - 0.033*ODI9 - 0.019*ODI10 + 0.002*age + 0.008*female + 0.018*education middle + 0.026* education high - 0.014*no partner - 0.079*secondary care - 0.064* tertiary care + 0.029*NRS moderate - 0.066*NRS severe

	Regression Coefficient (SE)	**95% CI**	
		2.5%	**97.5%**
Intercept	0.961 (0.025)	0.913	1.009
ODI1	-0.096 (0.003)	-0.101	-0.091
ODI2	-0.044 (0.002)	-0.048	-0.040
ODI3	-0.005 (0.002)	-0.009	-0.002
ODI4	-0.019 (0.002)	-0.023	-0.015
ODI5	-0.005 (0.002)	-0.008	-0.000
ODI6	-0.009 (0.002)	-0.012	-0.005
ODI7	-0.014 (0.002)	-0.018	-0.010
ODI9	-0.033 (0.002)	-0.037	-0.029
ODI10	-0.019 (0.002)	-0.023	-0.015
Age	0.002 (0.000)	0.001	0.002
Gender; female	0.008 (0.004)	0.000	0.016
Education; middle	0.018 (0.005)	0.009	0.027
Education; high	0.026 (0.005)	0.016	0.037
Partner; no partner	-0.014 (0.004)	-0.023	-0.006
Setting; secondary care	-0.079 (0.022)	-0.123	-0.036
Setting; tertiary care	-0.064 (0.022)	-0.107	-0.021
NRS; moderate	0.029 (0.008)	0.013	0.045
NRS; severe	-0.066 (0.009)	-0.083	-0.049

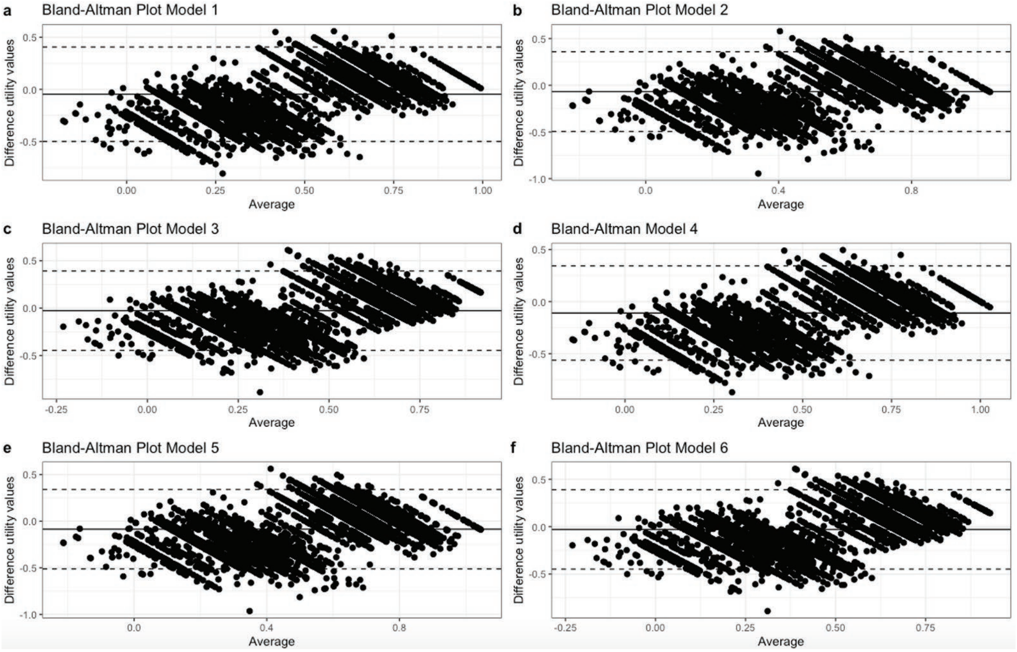
Model 6. Tobit with ODI individual items scores (ordered)

Utility = 0.831 - 0.006*ODI1;1 - 0.019*ODI1;2 - 0.153*ODI1;3 - 0.261*ODI1;4 - 0.262*ODI1;5 - 0.053*ODI2;1 - 0.098*ODI2;2 - 0.105*ODI2;3 - 0.190*ODI2;4 - 0.146*ODI2;5 - 0.002*ODI3;1 - 0.030*ODI3;2 - 0.009*ODI3;3 - 0.014*ODI3;4 - 0.041*ODI3;5 - 0.017*ODI4;1 - 0.033*ODI4;2 - 0.048*ODI4;3 - 0.069*ODI4;4 - 0.132*ODI4;5 - 0.004*ODI5;1 - 0.006*ODI5;2 - 0.009*ODI5;3 - 0.016*ODI5;4 -0.027*ODI5;5 - 0.002*ODI6;1 - 0.007*ODI6;2 - 0.013*ODI6;3 - 0.029*ODI6;4 - 0.045*ODI6;5- 0.003*ODI7;1 - 0.023*ODI7;2 - 0.036*ODI7;3 - 0.050*ODI7;4 - 0.051*ODI7;5 - 0.026*ODI9;1 - 0.036*ODI9;2 - 0.095*ODI9;3 - 0.155*ODI9;4 - 0.155*ODI9;5 -0.022*ODI10;1 - 0.043*ODi10;2 - 0.061*ODI10;3 - 0.080*ODI10;4 - 0.067*ODI10;5 + 0.0015134*age + 0.017 *education middle + 0.026* education high - 0.013*no partner - 0.099* secondary care - 0.089*tertiary care + 0.000 *NRS moderate - 0.081*NRS severe

	Regression Coefficient (SE)	95% CI	
		2.5%	97.5%
Intercept	0.831 (0.029)	0.774	0.888
ODI1;1	0.006 (0.018)	-0.029	0.042
ODI1;2	-0.019 (0.017)	-0.054	0.015
ODI1;3	-0.153 (0.018)	-0.188	-0.120
ODI1;4	-0.261 (0.018)	-0.297	-0.226
ODI1;5	-0.262 (0.023)	-0.306	-0.218
ODI2;1	-0.053 (0.005)	-0.063	-0.043
ODI2;2	-0.098 (0.006)	-0.109	-0.087
ODI2;3	-0.105 (0.008)	-0.122	-0.089
ODI2;4	-0.190 (0.016)	-0.220	-0.159
ODI2;5	- 0.146 (-0.040)	-0.224	-0.069
ODI3;1	-0.002 (0.010)	-0.021	0.018
ODI3;2	-0.030 (0.010)	-0.024	0.018
ODI3;3	-0.009 (0.010)	-0.028	0.011
ODI3;4	-0.014 (0.010)	-0.034	0.006
ODI3;5	-0.041 (0.013)	-0.066	-0.016
ODI4;1	-0.017 (0.005)	-0.027	-0.007
ODI4;2	-0.033 (0.006)	-0.045	-0.021
ODI4;3	-0.048 (0.007)	-0.063	-0.037
ODI4;4	-0.069 (0.010)	-0.088	-0.051
ODI4;5	-0.132 (0.028)	-0.187	-0.078
ODI5;1	-0.004 (0.008)	-0.012	0.020
ODI5;2	-0.006 (0.008)	-0.010	0.021
ODI5;3	-0.009 (0.009)	-0.026	0.008
ODI5;4	-0.016 (0.011)	-0.037	0.004
ODI5;5	-0.027 (0.018)	-0.061	0.008
ODI6;1	-0.002 (0.011)	-0.023	0.020
ODI6;2	-0.007 (0.011)	-0.029	0.015
ODI6;3	-0.013 (0.011)	-0.034	0.090
ODI6;4	-0.029 (0.011)	-0.047	-0.004
ODI6;5	-0.045 (0.014)	-0.071	-0.019
ODI7;1	-0.003 (0.006)	-0.015	0.008

	Regression Coefficient (SE)	95% CI	
		2.5%	97.5%
ODI7;3	-0.036 (-0.008)	-0.051	-0.021
ODI7;4	-0.050 (0.013)	-0.074	-0.025
ODI7;5	-0.051 (0.016)	-0.083	-0.020
ODI9;1	-0.026 (0.008)	-0.041	-0.011
ODI9;2	-0.036 (0.008)	-0.050	-0.021
ODI9;3	-0.095 (0.008)	-0.110	-0.080
ODI9;4	-0.155 (0.011)	-0.177	-0.133
ODI9;5	-0.155 (0.017)	-0.189	-0.121
ODI10;1	-0.022 (0.009)	-0.038	-0.005
ODI10;2	-0.043 (0.009)	-0.062	-0.025
ODI10;3	-0.061 (0.010)	-0.081	-0.041
ODI10;4	-0.080 (-0.011)	-0.103	-0.058
ODI10;5	-0.067 (0.012)	-0.091	-0.042
Age	0.002 (0.000)	0.001	0.002
Gender; female	-0.007 (0.004)	-0.001	0.015
Education; middle	0.017 (0.004)	0.008	0.025
Education; high	0.026 (0.005)	0.015	0.036
Partner; no partner	-0.013 (0.004)	-0.021	-0.004
Setting; secondary care	-0.099 (0.022)	-0.143	-0.056
Setting; tertiary care	-0.089 (0.022)	-0.132	-0.046
NRS; moderate	0.000 (0.008)	-0.016	0.016
NRS; severe	-0.081 (0.009)	-0.098	-0.064

Appendix C | Bland Altman plots estimated and actual utility values models 1–6 validation set



X-axis: Average measurement of the estimated and actual utility values, Y-axis: Difference in measurements between the two instruments.

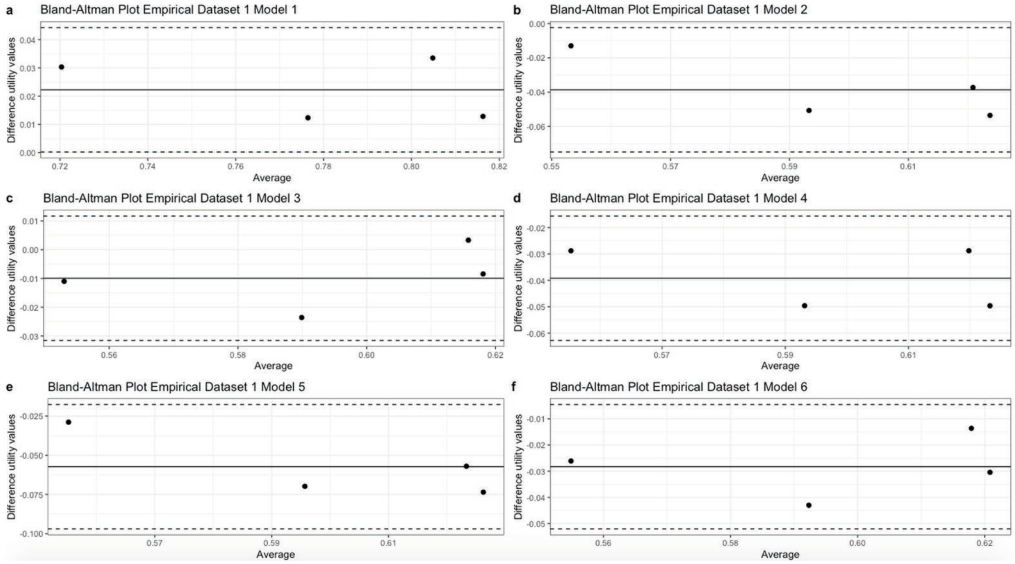
Solid line: Average difference in measurements between the estimated and actual utility values, Dashed lines: 95% confidence interval limits for the average difference.

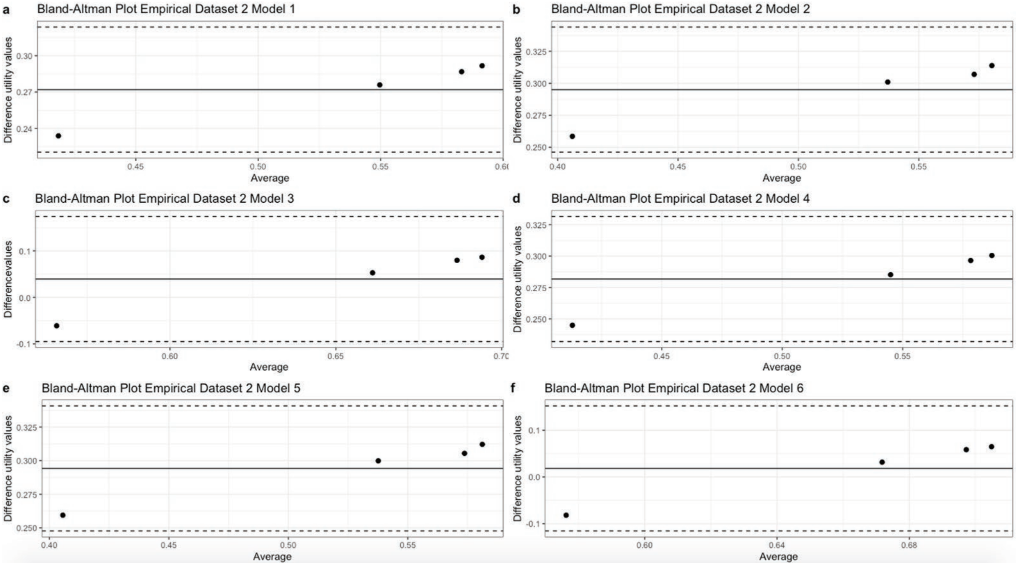
Appendix D | Sensitivity Analysis

SA 1 Mental Health	with mental health			without mental health		
	RMSE	R^2	AIC	RMSE	R^2	AIC
Model 1; OLS with ODI total scores	0.21	0.46	-738.80	0.22	0.43	-617.10
Model 2; OLS with ODI individual items scores (continuous)	0.20	0.49	-940.55	0.21	0.47	-802.74
Model 3; OLS with ODI individual items scores (ordered)	0.20	0.52	-1004.90	0.20	0.49	-869.50
Model 4; Tobit with ODI total scores	0.21	0.48	-699.13	0.22	0.44	-578.63
Model 5; Tobit with ODI individual items scores (continuous)	0.20	0.51	-904.42	0.21	0.48	-765.69
Model 6; Tobit with ODI individual items scores (ordered)	0.20	0.52	-960.76	0.20	0.49	-826.14
SA 2 Living with partner	**with variable partner**			**without variable partner**		
	RMSE	**R^2**	**AIC**	**RMSE**	**R^2**	**AIC**
Model 1; OLS with ODI total scores	0.22	0.45	-2326.48	0.22	0.45	-2318.62
Model 2; OLS with ODI individual items scores (continuous)	0.21	0.50	-3423.24	0.21	0.50	-3401.74
Model 3; OLS with Stepwise Selection AIC with ODI sub scores (ordered)	0.21	0.51	-3768.53	0.21	0.51	-3762.27
Model 4; Tobit with ODI total scores	0.22	0.46	-2054.46	0.22	0.46	-2061.91
Model 5; Tobit with Stepwise Selection AIC with ODI sub scores (continuous)	0.21	0.50	-3155.95	0.21	0.50	-3164.37
Model 6; Tobit with Stepwise Selection AIC with ODI sub scores (ordered)	0.21	0.51	-3467.56	0.21	0.51	-3473.60
SA3 Cross walk EQ-5D-3	**EQ-5D-3L**			**EQ-5D-5L reversed cross walk**		
	RMSE	**R^2**	**AIC**	**RMSE**	**R^2**	**AIC**
Model 1; OLS with ODI total scores	0.22	0.45	-2326.48	0.15	0.49	-12150.58
Model 2; OLS with ODI individual items scores (continuous)	0.21	0.50	-3423.24	0.15	0.53	-13158.25
Model 3; OLS with Stepwise Selection AIC with ODI sub scores (ordered)	0.21	0.51	-3769.51	0.14	0.54	-13412.33
Model 4; Tobit with ODI total scores	0.22	0.45	-2061.91	0.15	0.49	-12156.93
Model 5; Tobit with Stepwise Selection AIC with ODI sub scores (continuous)	0.21	0.50	-3164.37	0.15	0.53	-13158.25
Model 6; Tobit with Stepwise Selection AIC with ODI sub scores (ordered)	0.21	0.51	-3474.88	0.14	0.54	-13412.33

OLS: Ordinary Least Squares Regression, ODI: Oswestry Disability Index, RMSE: root-mean-square error, R2: proportion of variation in the dependent variable, AIC: Akaike information criterion

Appendix E | Bland Altman plots estimated and actual utility values models 1–6 empirical datasets





X-axis: Average measurement of the estimated and actual utility values, Y-axis: Difference in measurements between the two instruments.

Solid line: Average difference in measurements between the estimated and actual utility values, Dashed lines: 95% confidence interval limits for the average difference.

Appendix F | Regression coefficients per country for models 2 and 5

Regression coefficients Model 2

	UK	Spain	Japan	Zimbabwe	Germany	USA	South Korea	Denmark
Intercept	0.9216145	0.9629999	0.8061	0.8922	1.0250294	0.9124882	0.9344	0.8973506
ODI1	-0.1048651	-0.0912104	-0.03095	-0.05299	-0.0975522	-0.0648293	-0.03990	-0.0757004
ODI2	-0.0527377	-0.0623234	-0.02753	-0.03930	-0.0437269	-0.0386020	-0.03050	-0.0398935
ODI3	-0.0066237	-0.0083612	-0.003670	-0.004353	-0.0052832	-0.0044403	-0.004936	-0.0045352
ODI4	-0.0261759	-0.0311805	-0.01816	-0.01697	-0.0254695	-0.0176813	-0.02332	-0.0237435
ODI5	-0.0052813	-0.0059330	-0.004216	-0.003785	-0.0032939	-0.0043486	-0.004155	-0.0056333
ODI6	-0.0114667	-0.0124294	-0.005125	-0.006387	-0.0125239	-0.0070242	-0.006542	-0.0068254
ODI7	-0.0144788	-0.0125305	-0.005614	-0.008621	-0.0119149	-0.0095046	-0.006776	-0.0132994
ODI9	-0.0338144	-0.0360919	-0.01856	-0.02075	-0.0246979	-0.0230226	-0.02228	-0.0272122
ODI10	-0.0220272	-0.0247321	-0.01189	-0.01395	-0.0167978	-0.0154409	-0.01668	-0.0190554
Age	0.0017108	0.0016504	0.0007201	0.0009132	0.0012912	0.0011823	0.0008650	0.0013902
Sex; female	0.0082351*	0.0127427	0.009772	0.007933		0.0078843	0.01142	0.0095676
Education; middle	0.0162531	0.0158689	0.007196	0.009566	0.0112980	0.0111634	0.009091	0.0129300
Education; high	0.0218355	0.0219954	0.01072	0.01294	0.0141464	0.0148199	0.01210	0.0153532
No partner	-0.0113913	-0.0085004*	-0.004576	-0.005383	-0.0072071 *	-0.0088351	-0.003601	-0.0107631
Secondary care	-0.0820832	-0.0842040	-0.05519	-0.04963	-0.0788964	-0.0492312	-0.04828	-0.0481359
Tertiary care	-0.0656978	-0.0748289	-0.05581	-0.04609	-0.0620024	-0.0411599	-0.04427	-0.0364969
NRS; moderate	0.0388160	0.0359657	0.004944 *	0.01767	0.0395182	0.0239697	0.01842	0.0322145
NRS; severe	-0.0666840	-0.0567207	-0.01999	-0.03398	-0.0612173	-0.0416691	-0.01741	-0.0391295
R^2 model	0.5155	0.5427	0.5339	0.5631	0.4884	0.5289	0.5341	0.5173

* not significant

Regression coefficients Model 2

	France	Thailand	Canada	China	Italy	Singapore	Taiwan	Argentina
Intercept	0.8920794	0.8011386	0.8843	0.8854	0.9449	0.8237519	0.7539217	0.9289
ODI1	-0.0698340	-0.0586726	-0.05749	-0.04194	-0.05321	-0.0802764	-0.0730139	-0.05980
ODI2	-0.0734759	-0.0487556	-0.03653	-0.04289	-0.03633	-0.0687301	-0.0657758	-0.02262
ODI3	-0.0081591	-0.0058457	-0.003727	-0.005300	-0.004218	-0.0102091	-0.0081129	-0.004647
ODI4	-0.0279455	-0.0252811	-0.01788	-0.01989	-0.02106	-0.0302459	-0.0267910	-0.02291
ODI5	-0.0059446	-0.0047299	-0.004710	-0.003938	-0.005428	-0.0068003	-0.0055472	-0.004378
ODI6	-0.0131530	-0.0103183	-0.005809	-0.007129	-0.004731	-0.0153680	-0.0110460	-0.01036
ODI7	-0.0119208	-0.0090852	-0.01049	-0.007816	-0.009086	-0.0121587	-0.0120211	-0.007826
ODI9	-0.0366947	-0.0248885	-0.02213	-0.02582	-0.02105	-0.0482907	-0.0404785	-0.01948
ODI10	-0.0176721	-0.0153155	-0.01320	-0.01513	-0.01709	-0.0212139	-0.0224594	-0.01225
Age	0.0012969	0.0009576	0.001065	0.0008914	0.001171	0.0015746	0.0014902	0.0007053
Sex; female	0.0143939	0.0104631	0.007273	0.01227	0.009246	0.0151897	0.0177048	0.006471
Education; middle	0.0188068	0.0105884	0.01101	0.01318	0.007618	0.0255564	0.0215434	0.006118
Education; high	0.0325035	0.0162018	0.01475	0.02090	0.006850*	0.0457172	0.0349359	0.007121 *
No partner	-0.0100189	-0.0056038 *	-0.009769	-0.005817	-0.007207	-0.0126894	-0.0109666	
Secondary care	-0.1155225	-0.0892417	-0.04946	-0.06664	-0.02561 *	-0.1584970	-0.1102171	-0.09597
Tertiary care	-0.1209327	-0.0879129	-0.04343	-0.06882	-0.02004*	-0.1554560	-0.1116358	-0.07828
NRS; moderate	0.0080468 *	0.0131317	0.01803	0.008445	0.02737	-0.0012864*	0.0126286*	0.01409
NRS; severe	-0.0622104	-0.0425631	-0.03582	0.03084	-0.02272	-0.0773060	-0.0584446	-0.03883
R^2 Model	0.5757	0.5655	0.533	0.5753	0.515	0.557	0.5611	0.4742

* not significant

Regression coefficients Model 2

	Australia	Brazil	Chile	Hungary	Poland	Portugal	Sri Lanka	Sweden
Intercept	0.9025891	0.7976	0.8235944	0.9472674	0.9930241	0.7336	0.8370042	0.9092
ODI1	-0.0801515	-0.02961	-0.0632935	-0.0601439	-0.0920958	-0.04206	-0.0599403	-0.02571
ODI2	-0.0435985	-0.04156	-0.0591976	-0.0446507	-0.0406323	-0.05540	-0.0582604	-0.01638
ODI3	-0.0046462	-0.005133	-0.0081320	-0.0053581	-0.0055945	-0.007265	-0.0084620	-0.003047
ODI4	-0.0209143	-0.02042	-0.0271107	-0.0283114	-0.0253303	-0.02402	-0.0407823	-0.01112
ODI5	-0.0040522	-0.004258	-0.0058750	-0.0073975	-0.0052011	-0.006597	-0.0059588	-0.002437
ODI6	-0.0087226	-0.007792	-0.0099633	-0.0039895	-0.0089339	-0.009148	-0.0134058	-0.004826
ODI7	-0.0118253	-0.005723	-0.0108634	-0.0110644	-0.0127106	-0.007490	-0.0106433	-0.004880
ODI9	-0.0287889	-0.02157	-0.0375882	-0.0273724	-0.0271976	-0.02720	-0.0351063	-0.01726
ODI10	-0.0169747	-0.01159	-0.0238120	-0.0257401	-0.0217994	-0.01842	-0.0248097	-0.007187
Age	0.0013476	0.0006081	0.0013849	0.0015927	0.0014433	0.0009556	0.0012154	0.005130
Sex; female	0.0072436	0.01137	0.0170097	0.0144952	0.0063456*	0.01633	0.0191124	0.005607
Education; middle	0.0134860	0.009266	0.0180989	0.0098368	0.0104315	0.009909	0.0139406	0.009827
Education; high	0.0202586	0.01642	0.0289852	0.0064924*	0.0103657	0.01632	0.0208065	0.01747
No partner	-0.0110786	-0.003376*	-0.0083945	-0.0085097	-0.0064591*	-0.08909		-0.005084
Secondary care	-0.0705204	-0.08020	0.0893192		-0.0467255	-0.09395	-0.0996952	-0.05385
Tertiary care	-0.0601057	-0.08623	-0.0882755		-0.0273513*	0.003993*	-0.0989189	-0.05025
NRS; moderate	0.0255881	-0.002747*	0.0146376	0.0403118	0.0443564	0.005330	0.0213804	-0.0009108*
NRS; severe	-0.0541719	-0.02768	-0.0452285	-0.0159125	-0.0466466	-0.03400	-0.0330848	-0.02.275
R^2 model	0.53	0.5747	0.5681	0.4754	0.5109	0.5689	0.552	0.5174

* not significant

Regression coefficients Model 2

	Trinidad Tobago	Belgium	Finland	Iran	Malaysia	New Zealand	Slovenia
Intercept	0.8883515	0.8447444	0.7394	0.7589	0.9021951	0.8058	0.7929
ODI1	-0.0390160	-0.0675222	-0.01883	-0.03214	-0.0403614	-0.05954	-0.03393
ODI2	-0.0307738	-0.0396007	-0.03637	-0.06078	-0.0378891	-0.03396	-0.03881
ODI3	-0.0035665	-0.0051525	-0.003874	-0.006695	-0.0047368	-0.004221	-0.004225
ODI4	-0.0176425	-0.0182410	-0.01274	-0.01571	-0.0162989	-0.01650	-0.02606
ODI5	-0.0046815	-0.0031076	-0.003825	-0.005023	-0.0029200	-0.002557*	-0.002369
ODI6	-0.0048159	-0.0086009	-0.003055	-0.007477	-0.0069253	-0.007854	-0.01007
ODI7	-0.0071148	-0.0096681	-0.005288	-0.007503	-0.0060276	-0.008642	-0.007424
ODI9	-0.0154423	-0.0282179	-0.02356	-0.03182	-0.0224816	-0.02426	-0.02483
ODI10	-0.0122454	-0.0153905	-0.01171	-0.01354	-0.0123636	-0.01277	-0.01078
Age	0.0008015	0.0011642	0.0007940	0.0009362	0.0007950	0.0009665	0.0004864
Sex; female	0.0076287	0.0087419	0.01435	0.01686	0.0095461	0.007549	0.01211
Education; middle	0.0048550	0.0159546	0.01307	0.01732	0.0114116	0.01397	0.01465
Education; high	0.0045773*	0.0248290	0.02222	0.03269	0.0190117	0.02191	0.02579
No partner	-0.0042406	-0.0097974	-0.008083	-0.007903	-0.0053342	-0.008842	-0.006400
Secondary care	-0.0347116	-0.0818253	-0.05806	-0.09298	-0.0636079	-0.08000	-0.1092
Tertiary care	-0.0323348	-0.0744276	-0.06886	-0.01078	-0.0650069	-0.07378	-0.1153
NRS; moderate	0.0152583	0.0165661	-0.007442	-0.01283	0.0077327 *	0.01226	-0.003784*
NRS; severe	-0.0196063	-0.0503878	-0.02146	-0.04328	-0.0327250	-0.04561	-0.03126
R^2	0.5291	0.5195	0.471	0.5374	0.5658	0.5176	0.5395

* not significant

Regression coefficients Model 5

	UK	Spain	Japan	Zimbabwe	Germany	USA	South Korea	Denmark
Intercept 1	0.9472972	0.9895059	0.8120	0.9032	1.0534441	0.9276111	0.9454	0.9154778
Intercept 2	-1.4601898	-1.5052735	-2.332	-2.084	-1.5363619	-1.9010432	-2.118	-1.7364138
ODI1	-0.1060945	-0.0924513	-0.03121	-0.05350	-0.0988903	-0.0655513	-0.04043	-0.0765778
ODI2	-0.0524924	-0.0620764	-0.02748	-0.03920	-0.0434683	-0.0384583	-0.03037	-0.0397189
ODI3	-0.0068076	-0.0085525	-0.003711	-0.004431	-0.0054762	-0.0045482	-0.005017	-0.0046644
ODI4	-0.0261440	-0.0311491	-0.01815	-0.01696	-0.0254331	-0.0176635	-0.02340	-0.0237235
ODI5	-0.0056060	-0.0062683	-0.004289	-0.003924	-0.0036473*	-0.0045421	-0.004265	-0.0058678
ODI6	-0.0118429	-0.0128131	-0.005210	-0.006547	-0.0129204	-0.0072481	-0.006690	-0.0070971
ODI7	-0.0144099	-0.0124585	-0.00560	-0.008593	-0.0118338	-0.0094637	-0.006740	-0.0132512
ODI9	-0.0342032	-0.0364879	-0.01864	-0.02092	-0.0251148	-0.0232536	-0.02251	-0.0274933
ODI10	-0.0218918	-0.0245948	-0.01186	-0.01389	-0.0166507	-0.0153614	-0.01667	-0.0189594
Age	0.0016847	0.0016240	0.0007145	0.0009022	0.0012619	0.0011671	0.0008651	0.0013719
Sex; female	0.0085295	0.0130419	0.009837	0.008057		0.0080605	0.01135	0.0097820
Education; middle	0.0159807	0.0155870	0.007139	0.009451	0.0110166	0.0110042	0.008913	0.0127378
Education; high	0.0219722	0.0221218	0.01075	0.01299	0.0142850	0.0149028	0.01213	0.0154533
No partner	-0.0111703	-0.0082751*	-0.004524	-0.005286	-0.0069407*	-0.0087042		-0.0106022
Secondary care	-0.0960691	-0.0988243	-0.05854	-0.05567	-0.0944005	-0.0574566	-0.05516	-0.0579199
Tertiary care	-0.0792959	-0.0890656	-0.05907	-0.05197	-0.0771188	-0.0491595	-0.05105	-0.0460041
NRS; moderate	0.0336000	0.0306813	0.003827*	0.01548	0.0338432	0.0208932	0.01602	-0.0460041
NRS; severe	-0.0708414	-0.0609302	-0.02088	-0.03573	-0.0657410	-0.0441208	-0.01939	-0.0420809
R^2 model	0.5154651	0.5427006	0.533909	0.5630955	0.4882995	0.5288496	0.533964	0.5172503

* not significant

Regression coefficients Model 5

	France	Thailand	Canada	China	Italy	Singapore	Taiwan	Argentina
Intercept 1	0.9105936	0.8118602	0.8965	0.8959	0.9608	0.8432469	0.7693004	0.9428
Intercept 2	-1.6470017	-1.9013025	-1.997	-2.098	-1.973	-1.5005430	-1.6125055	-1.924
ODI1	-0.0706407	-0.0591457	-0.05807	-0.04240	-0.05398	-0.0811177	-0.0736964	-0.06042
ODI2	-0.0733158	-0.0486618	-0.03641	-0.04280	-0.03618	-0.0685617	-0.0656403	-0.02250
ODI3	-0.0082881	-0.0059205	-0.003641	-0.005374	-0.004334	-0.0103438	-0.0082196	-0.004745
ODI4	-0.0279223	-0.0252677	-0.01786	-0.01988	-0.02105	-0.0302196	-0.0267717	-0.02288
ODI5	-0.0061698	-0.0048615	-0.004866	-0.004068	-0.005639	-0.0070336	-0.0057366	-0.004546
ODI6	-0.0134090	-0.0104702	-0.005989	-0.007278	-0.004974	-0.0156361	-0.0112651	-0.01055
ODI7	-0.0118732	-0.0090583	-0.01045	-0.007789	-0.009042	-0.0121089	-0.0119813	-0.007793
ODI9	-0.0369600	-0.0250451	-0.02231	-0.02597	-0.02130	-0.0485695	-0.0407062	-0.01968
ODI10	-0.0175793	-0.0152620	-0.01314	-0.01508	-0.01700	-0.0211177	-0.0223827	-0.01218
Age	0.0012792	0.0009472	0.001053	0.0008813	0.001155	0.0015561	0.0014754	0.0006913
Sex; female	0.0145922	0.0105793	0.007414	0.01238	0.009436	0.0153978	0.0178762	0.006627
Education; middle	0.0186230	0.0104814	0.01088	0.01307	0.007445	0.0253681	0.0213923	0.005979*
Education; high	0.0325879	0.0162499	0.01482	0.02095	0.006931*	0.0458146	0.0350178	0.007173*
No partner	-0.0098760	-0.0055124*	-0.009664	-0.005731	-0.007059	-0.0125426	-0.0108399	
Secondary care	-0.1261291	-0.0953430	-0.05617	-0.07265	-0.03417	-0.1697453	-0.1189331	-0.1040
Tertiary care	-0.1312972	-0.0938662	-0.04996	-0.07468	-0.02836	-0.1664425	-0.1201396	-0.08606
NRS; moderate	0.0045179*	0.0110833	0.01556	0.006422*	0.02409	-0.0049764*	0.0096595*	0.01145
NRS; severe	-0.0650354	-0.0441987	-0.03779	-0.03246	-0.02532	-0.0802601	-0.0608164	-0.04093
R^2 model	0.5756387	0.565525	0.5329835	0.57526	0.51494	0.5570274	0.5610422	0.4741354

* not significant

Regression coefficients Model 5

	Australia	Brazil	Chile	Hungary	Poland	Portugal	Sri Lanka	Sweden
Intercept 1	0.9205291	0.8039	0.8390576	0.9594831	1.0188753	0.7424590	0.854156	0.9092
Intercept 2	-1.7247032	-2.244	-1.7118986	-1.6827221	-1.6065257	-1.9445409	-1.648065	-2.580
ODI1	-0.0810006	-0.02987	-0.0639867	-0.0613315	-0.0933810	-0.0424434	-0.060690	-0.02571
ODI2	-0.0434293	-0.04151	-0.0590597	-0.0443819	-0.0403758	-0.0553214	-0.058113	-0.01638
ODI3	-0.0047727	-0.005176	-0.0082410	-0.0055609	-0.0057844	-0.0073265	-0.008583	-0.003047
ODI4	-0.0208922	-0.02042	-0.0270919	-0.0283079	-0.0252997	-0.0240094	-0.040758	-0.01112
ODI5	-0.0042776	-0.004334	-0.0060682	-0.0077569	-0.0055373	-0.0067070	-0.006172	-0.002347
ODI6	-0.0089840	-0.007880	-0.0101857	-0.0043967	-0.0093227	-0.0092754	-0.013650	-0.004826
ODI7	-0.0117777	-0.005708	-0.0108226	-0.0110356	-0.0126399	-0.0074682	-0.010600	-0.004880
ODI9	-0.0290600	-0.02166	-0.0378188	-0.0277944	-0.0275985	-0.0273258	-0.035355	-0.01726
ODI10	-0.0168807	-0.01156	-0.0237348	-0.0256004	-0.0216604	-0.0183716	-0.024724	-0.007187
Age	0.0013295	0.0006022	0.0013700	0.0015644	0.0014163	0.0009470	0.001199	0.0005130
Sex; female	0.0074489	0.01144	0.0171836	0.0147854	0.0066487*	0.0164295	0.019308	0.005607
Education; middle	0.0132987	0.009206	0.0179436	0.0095554	0.0101440	0.0098232	0.013770	0.009827
Education; high	0.0203570	0.01645	0.0290658	0.0066940*	0.0104967	0.0163578		0.01742
No partner	-0.0109283	-0.003323*	-0.0082643	-0.0082528	-0.0062182*		0.020878	-0.005084
Secondary care	-0.0803519	-0.08387	-0.0980271		-0.0604633	-0.0941681	-0.109474	-0.05385
Tertiary care	-0.0696706	-0.08982	-0.0967684		-0.0406806*	-0.0989100	-0.108462	-0.05025
NRS; moderate	0.0219669	-0.003909*	0.0116307*	0.0353516	0.0389747	0.0023267*	0.018152	-0.0009108*
NRS; severe	-0.0570602	-0.02861	-0.0476263	-0.0198679	-0.0509219	-0.0353305	-0.035651	-0.02275
R^2 model	0.5300021	0.5746986	0.5680596	0.475328	0.510803	0.5688545	0.5519598	0.5174331

* not significant

Regression coefficients Model 5

	Trinidad Tobago	Belgium	Finland	Iran	Malaysia	New Zealand	Slovenia
Intercept 1	0.8972	0.8585886	0.7453	0.7684	0.9119	0.8161	0.8008
Intercept 2	-2.253	-1.8174829	-2.232	-1.943	-2.191	-1.952	-2.091
ODI1	-0.03942	-0.0681537	-0.01908	-0.03253	-0.04079	-0.06001	-0.03424
ODI2	-0.03069	-0.0394748	-0.03632	-0.06070	-0.03780	-0.03387	-0.03875
ODI3	-0.003629	-0.0052490	-0.003914	-0.006760	-0.004805	-0.004292	-0.004277
ODI4	-0.01763	-0.0182228	-0.01273	-0.01570	-0.01629	-0.01649	-0.02604
ODI5	-0.004795	-0.0032777	-0.003899	-0.005138	-0.003039	-0.002683*	-0.002459
ODI6	-0.004947	-0.0087993	-0.003142	-0.007610	-0.007062	-0.008001	-0.01018
ODI7	-0.007092	-0.009632	-0.005273	-0.007479	-0.006002	-0.008615	-0.007406
ODI9	-0.01558	-0.0284241	-0.02365	-0.03196	-0.02262	-0.02441	-0.02494
ODI10	-0.01220	-0.0153200	-0.01168	-0.01349	-0.01232	-0.01272	-0.01074
Age	0.0007929	0.0011506	0.0007885	0.0009273	0.0007857	0.0009564	0.0004793
Sex; female	0.007729	0.0088975	0.01442	0.01697	0.009653	0.007664	0.01219
Education; middle	0.004763	0.0158167	0.01301	0.01723	0.01132	0.01387	0.01458
Education; high	0.004618*	0.0249067	0.02225	0.03275	0.01906	0.02197	0.02583
No partner	-0.004159*	-0.0096840	-0.008030	-0.007824	-0.005256	-0.008757	-0.006341
Secondary care	-0.03957	-0.0895791	-0.06151	-0.09850	-0.06916	-0.08582	-0.1138
Tertiary care	-0.03706	-0.0819805	-0.07223	-0.1132	-0.07043	-0.07945	-0.1198
NRS; moderate	0.01351	0.0138492	-0.008571	-0.01461	0.005869*	0.01026*	-0.005173*
NRS; severe	-0.02100	-0.0525574	-0.02236	-0.04471	-0.03241	-0.04722	-0.03237
R^2 model	0.5290929	0.5194592	0.4709515	0.5374136	0.5657962	0.5176076	0.5394758

* not significant

CHAPTER 6

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

CHAPTER 7

The handling of missing data in trial-based economic evaluations: Should data be multiply imputed prior to longitudinal linear mixed-model analyses?

Ângela Jornada Ben, Johanna M. van Dongen, Mohamed El Alili, Martijn W. Heymans, Jos W. R. Twisk, Janet L. MacNeil-Vroomen, Maartje de Wit, Susan E. M. van Dijk, Teddy Oosterhuis , Judith E. Bosmans

European Journal of Health Economics, 2022

Abstract

Introduction

For the analysis of clinical effects, multiple imputation (MI) of missing data was shown to be unnecessary when using longitudinal linear mixed-models (LLM). It remains unclear whether this also applies to trial-based economic evaluations. Therefore, this study aimed to assess whether MI is required prior to LLM when analysing longitudinal cost and effect data.

Methods

Two-thousand complete datasets were simulated containing five time points. Incomplete datasets were generated with 10%, 25%, and 50% missing data in follow-up costs and effects, assuming a Missing At Random (MAR) mechanism. Six different strategies were compared using empirical bias (EB), root-mean-squared error (RMSE), and coverage rate (CR). These strategies were: LLM alone (LLM) and MI with LLM (MI-LLM), and, as reference strategies, mean imputation with LLM (M-LLM), seemingly unrelated regression alone (SUR-CCA), MI with SUR (MI-SUR), and mean imputation with SUR (M-SUR).

Results

For costs and effects, LLM, MI-LLM, and MI-SUR performed better than M-LLM, SUR-CCA, and M-SUR, with smaller EBs and RMSEs as well as CRs closers to nominal levels. However, even though LLM, MI-LLM and MI-SUR performed equally well for effects, MI-LLM and MI-SUR were found to perform better than LLM for costs at 10% and 25% missing data. At 50% missing data, all strategies resulted in relatively high EBs and RMSEs for costs.

Conclusion

LLM should be combined with MI when analysing trial-based economic evaluation data. MI-SUR is more efficient and can also be used, but then an average intervention effect over time cannot be estimated.

Introduction

Decisions about the implementation and/or reimbursement of new healthcare interventions increasingly rely on evidence of their cost-effectiveness compared to one or more alternative interventions, preferably usual care, to optimize the use of scarce healthcare resources.[1,2] Economic evaluations seek to provide this information by relating the difference in costs between healthcare interventions to the difference in effects.[2] In many cases, economic evaluations are performed alongside clinical trials, which are then referred to as trial-based economic evaluations.[3] Important advantages of trial-based economic evaluations include the prospective collection of cost and effect data as well as the use of patient-level information to draw inferences about the cost-effectiveness of healthcare interventions.[3]

Missing data are common in clinical trials as participants may skip questions, follow-up assessments, and/or drop out of the study.[4] In trial-based economic evaluations, costs are calculated as the sum of numerous cost components that are measured at different time points. Thus, if one cost component is missing, costs cannot be calculated.[2,3,5] In the literature, three different missing data mechanisms are distinguished that describe the association between the missing data and the observed and unobserved variables.[6] Missing Completely At Random (MCAR) assumes missing data to occur by chance. Hence, the missing data are not associated with observed or unobserved variables. Missing At Random (MAR) occurs when the missing data is associated with observed variables, but not with unobserved variables. Missing Not At Random (MNAR) occurs when missing data is associated with unobserved variables.[6]

Historically, missing data in trial-based economic evaluations were handled by simply deleting participants with missing values (i.e., complete-case analysis). However, deleting missing cases from the analysis potentially biases estimates, as systematic differences may exist between subjects with missing and complete data.[5–7] Other methods to account for missing data in trial-based economic evaluations include "naïve" imputation methods, such as mean imputation and last observation carried forward. The use of "naïve" imputation methods is discouraged, because these methods do not account for the uncertainty related to filling in the missing values like other more advanced methods do.[7–11] Advanced methods, such as Multiple Imputation (MI) and Longitudinal Linear Mixed-model analysis (LLM) with maximum likelihood estimation are therefore considered more valid and are increasingly used for handling missing data.[7–10,12,13]

The advantage of MI is that it estimates the total variance of the summary statistic considering the within- and between-imputation variance, reflecting the uncertainty around estimated missing values.[13] Another key advantage of MI is the possibility to include auxiliary variables in the imputation model that may be not relevant for the analysis model. This can help in improving the precision of the estimates and adjust for important missingness predictors. A possible disadvantage of such an approach, however, is the chance of mis-specifying the imputation model, which may in turn lead to incorrect results. Another possible disadvantage of MI is the added computational complexity, because outcomes need to be estimated for all imputed datasets an then they are pooled to obtain overall estimates. Also, MI may lead to unstable results when no sufficient imputations are performed.[14]

Previous studies have shown that when opting for LLM, in case of reasonably normally distributed outcomes, MI is not necessary to obtain unbiased estimates, because the maximum likelihood function uses all observed data and produces unbiased estimates under the MAR assumption.[14,15] Faria et al. (2014) suggested that MI is not required prior to LLM in trial-based economic evaluations either, but this has never been empirically tested.[5] This is important, however, because there are three distinct statistical challenges to trial-based economic evaluations that may affect the performance of LLM to deal with missing data: 1) costs are typically heavily right-skewed; 2) costs are cumulative sums over time, and 3) costs and effects are correlated[15]. This study aimed to bridge this gap in knowledge by assessing whether MI of missing variables prior to LLM increases its performance when analysing longitudinal cost and effect data. Additionally, the impact of using either one of these approaches on cost-effectiveness estimates was assessed using empirical data.

Methods

To assess whether MI of missing values is required prior to LLM, a simulation study was conducted. SUR and mean imputation were added as analytic strategies to assess the relative performance of the main models (i.e., LLM and MI-LLM) versus (simpler) alternative approaches. Mean imputation was added because the results of Sullivan et al. (2016) suggest that MI is not the only acceptable way to handle missing data in RCTs and that simpler approaches, such as mean imputation, may also have satisfactory performance. SUR was added, because with LLM the possible correlation between costs and effects is neglected, whereas SUR can account for this correlation through correlated error terms.[16,17] In total, we compared six different methodological strategies; (i) LLM alone (LLM), (ii) mean imputation combined with LLM (M-LLM), (iii) MI combined with LLM (MI-LLM), (iv) seemingly unrelated regression alone (SUR-CCA), (v) mean imputation combined with SUR (M-SUR), and (vi) MI combined with SUR (MI-SUR). Additionally, two empirical datasets were used to evaluate the external validity of the results.[18,19]

Simulated datasets

Complete data generation

Two thousand complete datasets were generated using the *Simstudy* R package (R statistical software − version 3.5.2).[20] The parameters used for generating variables were based on previous trial-based economic evaluations[21,22] and the empirical datasets,[18,19] and are summarized in Table 1. The R code for data generation can be found in Supplementary material 1.

A total of 600 subjects per dataset was generated. An intervention to control ratio of 52:48 was simulated to resemble slightly unbalanced empirical datasets using a binomial distribution. A complete set of baseline variables was generated, including age, gender, costs (cT0), and utility values (uT0; i.e. a measure of health-related quality of life, HRQoL).[23] Age was generated using a normal distribution, gender using a binomial distribution, while costs and utility values were generated with a gamma distribution (Table 1).[24] At baseline, age and gender were positively related with costs and utility values, indicating that we assumed older subjects and men to have higher costs and utility values than younger subjects and women, respectively. To create a plausible MAR assumption, age and gender were slightly imbalanced, but we made sure that these imbalances were small and in line with those encountered in previous trial-based economic evaluations.[19,21,22,25] Additionally, we assumed the correlation between costs and utility values to be about -0.50.[26,27] Such a negative correlation might appear when subjects with a lower health-related quality of life (i.e. lower utility values) have higher treatment costs, because they were less healthy to begin with.[27]

Follow-up cost and utility values were simulated, with a true total cost and quality-adjusted life-year (QALY) difference during the complete-duration of follow-up of 250 euros and 0.04 QALY, respectively (Table 1, Supplementary Material 1). This was done by first generating costs and utility values at 3-month follow-up (i.e., cT1 and uT1), based on the subjects' baseline cost and utility values (i.e., cT0 and uT0) as well as their age, gender, and treatment allocation (trt). Then, all other follow-up cost and utility values (i.e., 6, 9, and 12-month follow-up) were extrapolated from the subjects' 3-month follow-up values, using a correlation between time points of 0.7 for costs and 0.9 for utility values and a compound symmetry correlation structure (i.e., correlations between subsequent time-points were assumed to be the same).

Table 1 | Parameter values

Parameters	Value	Motive
Age	Mean of 40 years with a 5-year difference between treatment groups (trt) and a variance of 140. Data were simulated using a normal distribution and a minimum of 18 and maximum of 99 years. *Formula:* *age = "40 + (5*trt)", variance = 140, dist="normal"* *data$age <= 18, 18, data$age* *data$age >= 99, 99, data$age*	Mean age and distribution were in line with those that might be observed in empirical trial-based economic evaluations. The variance was tweaked up until the point of getting a standard deviation similar to that encountered in empirical data.[18,19,21,22] A slight baseline imbalance was simulated to enable the generation of a plausible MAR assumption.
Gender	The mean proportion of male subjects was 52% and the proportion of female subjects 48%. Data were simulated using a binary distribution. *Formula:* *Gender = "0.52 + (0.4*trt)", dist = "binary"*	The proportion of male subjects was in line with those that might be observed in empirical trial-based economic evaluations.[18,19,21,22] A slight baseline imbalance was simulated to enable the generation of a plausible MAR assumption.
Utility values at baseline (uT0)	uT0 was generated dependent on age and gender, with a variance of 0.002, and a gamma distribution with logit link function. *Formula:* *uT0 = "0.2 + (0.0045 * age) + (0.025 * gender)", variance = 0.002, dist = "gamma", link = "logit"*	The dependency on age and gender allowed for the generation of a plausible MAR assumption. A gamma distribution with a logit link function and a variance of 0.002 were used as these parameters allowed for the generation of utility values ranging from 0 to 1. In doing so, we made sure that the average values and variances were in line with those that might be encountered in empirical trial-based economic evaluations.[18,19,21,22]

Parameters	Value	Motive
Costs at baseline (cT0)	Mean cT0 was generated dependent on age and gender, with a variance of 0.15, and a gamma distribution with logit link function. _Formula:_ _“50 + (15*age) + (100*gender)”, variance = 0.15, dist = “gamma”, link = “logit”_ cT0 and uT0 were generated with a negative correlation of -0.5 _data <- addCorFlex(data, defb, rho = -0.5)_	The dependency on age and gender allowed for the generation of a plausible MAR assumption. A gamma distribution with a logit link function and a variance of 0.15 were used to generate baseline costs similar to those that might be encountered in empirical trial-based economic evaluations.[18,19,21,22] A negative correlation between costs and utilities was set, as higher costs are generally associated with lower health-related quality of life.[18,19,21,22]

Utility values at T1 (uT1) and follow-up utilities	Mean uT1 was generated dependent on uT0, age, and gender with a difference of 0.04 between treatment groups and a variance of 0.002 following a gamma distribution with logit link function.
	Formula:
	_uT1 = “uT0 + (0.04* trt) + (0.002 * age) + (0.01 * gender)”, variance = 0.002, dist = “gamma”, link = “logit”_
	Utilities at time points T2, T3, and T4 were generated based on uT1 with a correlation of 0.9 and a variance of 0.005.
	Q <- matrix(c(1.0,0.9,0.9,0.9,1.0,0.9,0.9,0.9,1.0), nrow = 3)
	data <- addCorGen(dtOld = data, idvar = “id”, nvars = 3, corMatrix = Q, dist = “normal”, param1 = “uT1”, param2 = “variance”, cnames = “uT2, uT3, uT4”)
		A difference in follow-up utility values of 0.04 was generated per time point based on the minimally important difference for this outcome that has previously been found to range from 0.03 to 0.52.[54,55]
		A gamma distribution with a logit link function and a variance of 0.002 were used to generate follow-up utilities similar to those that might be encountered in empirical trial-based economic evaluations.[18,19,21,22]

Costs at T1 (cT1) and follow-up costs	Mean cT1 was generated dependent on cT0, age, and gender with a difference of €62.5 between treatment groups per time point and a variance of 0.15 following a gamma distribution with logit link function. _Formula:_ _“cT0 + (62.5*trt) + (1*age) + (10*gender)”, variance = 0.15, dist = “gamma”, link = “logit”_ Costs at time points T2, T3, and T4 were generated based on cT1 with a correlation of 0.7 and a variance of 0.1.	A difference in follow-up costs of €62.5 was simulated per time point to generate a total cost difference during follow-up of €250. A gamma distribution with a logit link function and a variance of 0.15 were used to generate follow-up costs similar to those that might be encountered in empirical trial-based economic evaluations.[18,19,21,22]

Missing data generation

We assumed baseline data to be completely observed for all subjects, which is often the case in trial-based economic evaluations.[19,21,22,25] Missing follow-up cost and utility values were generated under the assumption that the missingness of data was solely due to drop-out, resulting in monotone missing data patterns only (Supplementary Material 2). This means that subjects were assumed to either complete the study, and hence all follow-up assessments, or to drop-out after 3, 6, 9, or 12-months and have missing data from that time point on. We acknowledge that intermittent missingness patterns (e.g., non-monotone) may also exist in empirical data. However, for the sake of simplicity, and considering previous literature that shows that drop-out is frequent for EQ-5D data,[28,29] we opted to simulate monotone missing data patterns only.

To generate follow-up missing values, a missing data indicator m_{ij} was generated for every subject *i* (i = 1, …, N=600) at time point *j* (j = 1, …, 4) using a binomial distribution and a logit link function to model the linear dependence between the probability of missing data and its predictive variables (i.e., age, gender, cT0, uT0, trt).

$$m_{ij} \sim Binomial(\pi_{mij}),$$

$$logit(\pi_{mij}) = \beta_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 uT0_i + \beta_4 cT0_i + \beta_5 trt_i + \varepsilon_i, \tag{1}$$

where π_{mij} is the probability of a subject i having missing data at time point *j*. The intercept (β_0) and coefficients ($\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$) of covariates were tweaked to generate a plausible MAR assumption and datasets with 10%, 25%, and 50% missings. ε_i is the error term (Supplementary material 2).[28] Then, at the respective time point *j*, all complete follow-up cost and utility values were replaced by missing values according to m_{ij}. This process was repeated for all consecutive time points amongst the subset of individuals still in the study.[28] Thus, the "missingness" of costs and utility values at all time points was conditional on age, gender, as well as their baseline values. Subjects in the intervention group were assumed to be 4, 2, and 1 times more likely to have missing values compared to their control group counterparts for the datasets with 10%, 25%, and 50% missing data, respectively. This was done to strengthen the MAR assumption. The MAR assumption in our simulation differs from the MNAR assumption, because the m_{ij} was independent of the partially observed follow-up cost and utility values.[6,7]

Number of simulations

The number of required simulated data sets (n_{sim}) was calculated using the Monte Carlo standard error of the expected coverage rate.[30] The Monte Carlo standard error quantifies simulation uncertainty and provides an estimate of the standard error for simulation performance measures when using a finite number of simulations.[30] The coverage rate of the confidence intervals, that uses information of the Monte Carlo standard error represents the probability that a confidence interval contains the 'true' value. In order to estimate the coverage rate with an acceptable degree of imprecision, a total of 1900 simulated datasets (rounded up to 2000) was found to be needed based on a maximal Monte Carlo standard error of 0.5 and an expected coverage rate of 95%.

Data analyses

Data analyses were performed in StataSE 16® (StataCorp LP, CollegeStation, TX, US). All Stata codes can be found in Supplementary material 2, 3, and 4.

Methodological strategies

Longitudinal Linear Mixed-model analysis (LLM) – Two separate LLMs were performed, including one for costs and one for utility values:

$$Costs_{ij} = \beta_{1c}time_j + \beta_{2c}trt_i + \beta_{3c}time_j trt_i + \beta_{4c}cT0_i + \beta_{5c}time_j cT0_i + \beta_{6c}age_i + \beta_{7c}time_j age_i + \beta_{8c}gender_i + \beta_{9c}time_j gender_i + \omega_{ci} + \varepsilon_{cij},$$

$$Utility_{ij} = \beta_{1u}time_j + \beta_{2u}trt_i + \beta_{3u}time_j trt_i + \beta_{4u}uT0_i + \beta_{5u}time_j uT0_i + \beta_{6u}age_i + \beta_{7u}time_j age_i + \beta_{8u}gender_i + \beta_{9u}time_j gender_i + \omega_{ui} + \varepsilon_{uij},$$

$$\omega_i \sim Normal(0, \sigma_\omega^2),\ \varepsilon_{ij} \sim Normal(0, \sigma_\varepsilon^2) \qquad (2)$$

where $Costs_{ij}$ and $Utility_{ij}$ represent the cost and utility values of subject i (i = 1, …, N=600) at time point j (j = 1, …, 4). The model parameters include the coefficients $\beta_{1,...,}\beta_9$ of covariates, including various – by $time_j$ interactions. Moreover, ω_{ci} and ω_{ui} represent the random intercepts and ε_{cij} and ε_{cij} the error terms at each time point j for costs and utility, respectively. Both ω_i and ε_{ij} follow a normal distribution.[28,31,32]

To calculate the total cost difference between treatment groups, information was extracted on the average cost differences per time point, after which all follow-up cost differences were summed. To estimate the difference in QALY between treatment groups, information was extracted on the average utility differences per time point, after which the area under the curve method was applied.[23] Detailed information on model specification can be found in Supplementary material 3.

Mean Imputation combined with LLM (M-LLM) – In this strategy, missing cost and utility values were replaced by the mean values from the available cases at each time point (i.e., unconditional mean imputation).[5] Subsequently, two separate LLMs were fitted as outlined under LLM (Supplementary material 3).

Multiple Imputation combined with LLM (MI-LLM) – In this strategy, missing cost and utility values were first imputed using Multivariate Imputation by Chained Equations (MICE; FCS-standard)[33] with Predictive Mean Matching (PMM).[34] With PMM, a case with one or more missing values is matched with a number of cases with complete data (i.e., donor observations).[34] The characteristics used to match an observation with missing values with donor observations are the variables specified in the imputation model, which in our case included age, gender, cT0, and uT0. Subsequently, a value is randomly drawn from the donor observations that have an observed value for that variable.[34] The MICE algorithm then uses this random value to fill in missing data in an iterative process until the pre-specified imputation model converges.[33] A set of 5 donors with complete data was used for the matching (i.e., a k-nearest neighbour [knn] of 5).[34] The imputation model was stratified by treatment group (i.e., trt).[17] In total, 10 datasets were imputed for datasets with 10% and 25% missing data, and 20 for datasets with 50% missing data. This was done to ensure that the loss-of-efficiency was smaller than 0.05.[33] Subsequently, two separate LLMs were fitted per imputed dataset as outlined under LLM, after which pooled estimates were obtained using Rubin's rules[6] (Supplementary material 3).

Seemingly Unrelated Regressions − Complete Case Analysis (SUR-CCA) – In this strategy, total costs and QALY were calculated by adding costs at each time point and using the area under the curve method, respectively.[23] Subsequently, a seemingly unrelated regressions (SUR) model was fitted. A SUR model consists of two separate regression equations, e.g. one for total costs and one for QALY, while simultaneously correcting for their possible correlation through correlated error terms.[16,35] With the current strategy, only subjects with completely observed data were analysed (i.e., a complete-case analysis).

$$Costs_i = \beta_{0c} + \beta_{1c}trt_i + \beta_{2c}cT0_i + \beta_{3c}age_i + \beta_{4c}gender_i + \varepsilon_{ci}$$

$$QALY_i = \beta_{0q} + \beta_{1q}trt_i + \beta_{2q}uT0_i + \beta_{3q}age_i + \beta_{4q}gender_i + \varepsilon_{qi}$$

$$\begin{pmatrix}\varepsilon_{ci}\\ \varepsilon_{qi}\end{pmatrix} \sim Normal\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_c^2 & \sigma_{cq}\\ \sigma_{cq} & \sigma_q^2\end{pmatrix}\right) \tag{3}$$

Where $Costs_i$ and $QALY_i$ and are the observed total costs and QALY during follow-up of subject i ($i = 1, \ldots, N$). β_{0c} and β_{0e} represent the models' intercept, β_{1c} and β_{1q} represent the regression coefficients of the independent variable 'treatment group' (trt). β_{2c} ... β_{4c} and β_{2q} ... β_{4q} represent the regression coefficients for baseline values, age, and gender, ε_{ci} and ε_{ei} represent the correlated error terms for costs and QALY, respectively.[16,35]

Mean Imputation combined with SUR (M-SUR) – In this strategy, follow-up missing cost and utility data were replaced by the mean values from the available cases at each time point (i.e., unconditional mean imputation).[5] Subsequently, total costs and QALYs were calculated, and SUR analyses were performed as outlined under SUR (Supplementary material 3).

Multiple Imputation combined with SUR (MI-SUR) – In this strategy, follow-up missing cost and utility data were first imputed using Multivariate Imputation by Chained Equations (MICE) as outlined under MI-LLM.[32] Then, total costs and QALYs were calculated, and a SUR was fitted per imputed dataset as outlined under SUR, after which pooled estimates were obtained using Rubin's rules (Supplementary material 3).[32]

Comparison of the methodological strategies

The performance of the methodological strategies was assessed with regard to costs and QALY differences using the following performance measures: empirical bias (EB), root-mean-square error (RMSE) and coverage rate (CR).[30] Performance measures were calculated by comparing the 'true' values (i.e., estimand, [θ]) with the estimated values ($\hat{\theta}$) obtained from the methodological strategies conducted over the 2000 simulated datasets. Monte Carlo standard errors were estimated for each performance measure to quantify the uncertainty of these measures due to using 2000 simulated datasets[30] (Supplementary material 3).

(i) Empirical bias (EB) represents the average difference between $\hat{\theta}$and θ. This performance measure estimates whether a method targets θ on average and must, therefore, be small.

$$EB = \frac{1}{n_{sims}} \sum_{i=1}^{n_{sims}} \left(\widehat{\theta_\imath} - \theta \right) \tag{4}$$

(ii) Root-mean-square error (RMSE) represents the square root of the difference between $\hat{\theta}$and θ.

$$RMSE = \sqrt{\frac{1}{n_{sims}} \sum_{i=1}^{n_{sims}} \left(\widehat{\theta_\imath} - \theta\right)^2} \tag{5}$$

The mean-square error (MSE) is a measure of accuracy that combines the bias and variance in a single measure.[35] For easier interpretation, we report the square root of the MSE, to express it on the same scale as costs and QALYs.[35]

(iii) Coverage rate (CR) represents the percentage of times that the θ is covered by the estimated 95% CI of $\hat{\theta}$.

$$CR = \frac{1}{n_{sims}} \sum_{i=1}^{n_{sims}} 1\left(\widehat{CB_{lower,\imath}} \leq \theta \ \leq \widehat{CB_{upper,\imath}}\right) \tag{6}$$

The Monte Carlo standard error distance from the nominal value of 0.95 was used as a criterion of poor coverage.[36] It is worth to note that CR alone may mislead conclusions about the accuracy of methodological strategies, because high variances in estimates can lead to high CR.[30,37] Therefore, in this study, CR was evaluated jointly with the other two performance measures.[30]

In addition to assessing the performance of the methodological strategies for handling missing data (see section 2.2.1), we assessed the performance of LLM and SUR in the complete datasets. This was done to have a better understanding of their performance in the context of a trial-based economic evaluation in general. That is, before examining their performance for handling missing data).

Empirical datasets

Description of the datasets

Data from two pragmatic randomized controlled trials were used in addition to the simulated data. In the first trial (empirical dataset 1), the cost-effectiveness of early rehabilitation after lumbar disc surgery was compared to no referral.[18] For the current study, utility values collected at baseline, 12, and 26 weeks and costs collected at 6, 12, and 26 weeks were used. For all scenarios, mean imputation was used to impute missing values at baseline.[5] Of the 169 participants used in our study, 13% (n=22) had missing cost and/or utility data at one or more follow-up time points (Supplementary Material 4). Stepwise backwards regression models with p<0.05, were used to identify baseline variables that were predictive of the missingness of data and/or the cost-effectiveness outcomes. The identified variables were added to the imputation model as auxiliary variables (i.e., age, level of education, utility values, Oswestry Disability Index [ODI], pain intensity, Örebro Musculoskeletal Pain Screening Questionnaire [OMPSQ], and the credibility and expectancy surgery [CEQ]).[18] Missing cost and utility data were imputed using Multivariate Imputation by Chained Equations (MICE; FCS-standard)[33] with PMM, stratified by treatment group.[34] Ten datasets were imputed to guarantee a loss of efficiency <0.05. SUR and LLM were then fitted to the imputed data. The LMM models (i.e., M-LLM and MI-LLM) and SUR models (i.e., SUR-CCA, M-SUR, and MI-SUR) did not include auxiliary variables, and only were corrected for confounders (i.e., baseline utility values, ODI, OMPSQ, and CEQ). The LLM analysis model included both, auxiliary variables and confounders as, in doing so, it should lead to similar results when compared to MI-LLM[5] (Supplementary material 4).

In the second trial (empirical dataset 2), the cost-effectiveness of an interpersonal psychotherapy for older adults with major depression was compared to care as usual (i.e., control). For this study, utility values collected at baseline, 6, and 12 months and costs collected at 2, 6, and 12 months were used.[19] Mean imputation was used to impute missing values at baseline.[5] Of the 143 participants, 68% (n=98) of cost and utility data were missing at one or more follow-up time points (Supplementary material 4). Stepwise backwards regression models with p<0.05, were used to identify baseline variables that were predictive of the missingness of data and/or the cost-effectiveness outcomes. The identified variables were added to the imputation model as auxiliary variables (i.e., age, activity daily living [ADL], utility values, alcohol-induced disorder, and mental health problems utility values, marital status, and household composition).[19] Missing cost and utility data were imputed using Multivariate Imputation by Chained Equations (MICE; FCS-standard)[33] with PMM by treatment group.[34] Twenty datasets were imputed to guarantee a loss of efficiency <0.05. SUR and LLM were then fitted to the imputed data. The LMM models (i.e., M-LLM and MI-LLM) and SUR models (i.e., SUR-CCA, M-SUR, and MI-SUR) did not include auxiliary variables and only were corrected for confounders (i.e., baseline utility values, marital status, and household composition). The LLM analysis model included both, auxiliary variables and confounders[5] (Supplementary material 4).

Cost-effectiveness analysis of the empirical datasets

Cost-effectiveness analyses were performed using the differences in costs and QALY between treatment groups estimated by all of the methodological strategies described under 2.2.1. to both empirical datasets. Cost-effectiveness acceptability curves (CEACs) were estimated using the parametric Incremental Net Benefit (INB) approach, where the estimate of acceptability was obtained as the probability that INB>0 for every value of the willingness-to-pay threshold (WTP).[39,40] The INB is defined as $INB = \lambda \text{ x } \Delta_q - \Delta_c$ and the probability (Pr) of the INB being positive conditional to λ is estimated as:

$$Pr(INB > 0|\lambda) = 1 - \Phi\left(\frac{\lambda \times \Delta_q - \Delta_c}{\sqrt{(\lambda^2 \times Var(\Delta_q) + Var(\Delta_c) - 2 \times \lambda \times Cov(\Delta_q, \Delta_c)}}\right) \quad (7)$$

Where λ is the WTP, Φ is the cumulative standard normal distribution, and Δ_q, Δ_c are the differences in QALY and total costs between the intervention and control, respectively. $Var(\Delta_q)$, $Var(\Delta_c)$,are the variances around differences in QALY and total costs, and $Cov(\Delta_q, \Delta_c)$ is the covariance. To facilitate interpretation of the results, the probability of cost-effectiveness was also reported for willingness-to-pay thresholds of 10,000, 20,000 and 50,000 € per QALY gained.[41]

Results

Comparison of methodological strategies

Simulated datasets

For costs, LLM, MI-LLM and MI-SUR resulted in lower EBs and RMSEs compared to M-LLM, SUR-CCA, and M-SUR for all proportions of missing data. For 10% and 25% of missing data, MI-LLM and MI-SUR resulted in lower EBs and RMSEs compared to LLM. Moreover, LLM and MI-LLM were associated with relatively high levels of overcoverage for 10% and 25% of missing data, whereas for MI-SUR CR was closest to the nominal value. For 50% of missing data, all methodological strategies resulted in relatively higher EBs and RMSEs compared to those found in 10% and 25% of missing data. For LLM, CR were closest to the nominal value compared to the other methodological strategies at 50% of missing data.

For QALY, LLM, MI-LLM, and MI-SUR resulted in lower EBs and RMSEs compared to M-LLM, SUR-CCA, and M-SUR for all proportions of missing data. For LLM, EBs and RMSEs were similar or slightly lower than for MI-LLM and MI-SUR for all proportions of missing data. For 10% and 25% of missing data, LLM and MI-LLM were associated with small levels of overcoverage, while MI-SUR was associated with small levels of undercoverage. For 50% of missing data, LLM, MI-LLM, and MI-SUR presented small levels of undercoverage (Table 2, Figure 1). For both, costs and QALY outcomes, LLM, MI-LLM, and MI-SUR has similar performance measure compared to the LLM and SUR models based on complete datasets.

Table 2 | Performance measures of the methodological strategies for Costs and QALY

	Costs, €	LLM	M-LLM	MI-LLM	SUR-CCA	M-SUR	MI-SUR
Complete data	EB (MCse)	0.19 (3)	NA	NA	0.17 (3)	NA	NA
	RMSE (MCse)	153 (27)	NA	NA	153 (27)	NA	NA
	CR (MCse)	0.995 (0.158)	NA	NA	0.957 (0.456)	NA	NA
Missing 10%	EB (MCse)	-30 (4)	-92 (4)	-6 (3)	-106 (4)	-92 (4)	-6 (3)
	RMSE (MCse)	164 (29)	184 (32)	157 (27)	199 (34)	184 (32)	157 (27)
	CR (MCse)	0.993 (0.180)	0.981 (0.305)	0.993 (0.180)	0.913 (0.630)	0.895 (0.685)	0.948 (0.496)
Missing 25%	EB (MCse)	-46 (4)	-97 (4)	-10 (3)	-163 (6)	-97 (4)	-10 (3)
	RMSE (MCse)	182 (32)	195 (34)	157 (27)	269 (47)	195 (34)	157 (27)
	CR (MCse)	0.987 (0.253)	0.966 (0.402)	0.996 (0.141)	0.892 (0.694)	0.855 (0.787)	0.951 (0.483)
Missing 50%	EB (MCse)	-43 (5)	-224 (7)	-38 (5)	-167 (6)	-224 (7)	-38 (5)
	RMSE (MCse)	223 (40)	333 (56)	214 (38)	285 (49)	333 (56)	214 (38)
	CR (MCse)	0.950 (0.485)	0.671 (1.050)	0.972 (0.368)	0.860 (0.776)	0.507 (1.118)	0.925 (0.589)
	QALY	**LLM**	**M-LLM**	**MI-LLM**	**SUR-CCA**	**M-SUR**	**MI-SUR**
Complete data	EB (MCse)	-0.0000589 (0.0001004)	NA	NA	-0.0000586 (.0001004)	NA	NA
	RMSE (MCse)	.0044895 (.0008193)	NA	NA	0.0044890 (0.0008194)	NA	NA
	CR (MCse)	0.914 (0.625)	NA	NA	0.948 (0.496)	NA	NA
Missing 10%	EB (MCse)	-0.0000687 (0.0001050)	-0.0043361 (0.0001455)	-0.0001805 (0.0001053)	0.0010317 (0.0001113)	-0.0043359 (0.0001455)	-0.0001803 (0.0001053)
	RMSE (MCse)	0.0046931 (0.0008550)	0.0065054 (0.0011015)	0.0047081 (0.0008567)	0.0049750 (0.0009084)	0.0065048 (0.0011014)	0.0047078 (0.0008568)
	CR (MCse)	0.975 (0.346)	0.906 (0.651)	976 (0.342)	0.951 (0.483)	0.847 (0.804)	0.925 (0.476)
Missing 25%	EB (MCse)	-0.0000731 (0.0001151)	-0.0026774 (0.0001314)	-0.0000688 (0.0001071)	0.0016466 (0.0001443)	-0.0026772 (0.0001314)	-0.0000684 (0.0001071)
	RMSE (MCse)	0.0051474 (0.0009316)	0.0058771 (0.0010395)	0.0047891 (0.0008516)	0.0064497 (0.0011593)	0.0058766 (0.0010395)	0.0047889 (0.0008516)
	CR (MCse)	0.966 (0.402)	0.934 (0.555)	0.978 (0.328)	0.950 (0.487)	0.860 (0.774)	0.944 (0.514)
Missing 50%	EB (MCse)	-0.0001218 (0.0001422)	-0.0220931 (0.0005334)	-0.0017733 (0.0001462)	0.00167082 (0.0001542)	-0.0220928 (0.0005334)	-0.00177395 (0.0001462)
	RMSE (MCse)	0.0063599 (0.0011411)	0.0238472 (0.0030351)	0.0065377 (0.0011707)	0.0068928 (0.0012258)	0.0238469 (0.0030351)	0.00653833 (0.0011710)
	CR (MCse)	0.936 (0.545)	0.095 (0.657)	0.947 (0.501	0.932 (0.561)	0.085 (0.625)	0.926 (0.583)

LLM: longitudinal mixed-model. M-LLM: Mean imputation combined with LLM. MI-LLM: Multiple imputation combined with LLM. SUR-CCA: Seemingly unrelated regressions - complete case analysis. M-SUR: mean imputation combined with SUR. MI-SUR: multiple imputation combined with SUR. MCse: Monte Carlo standard error. EB: empirical bias. RMSE: root-mean-square error. CR: coverage rate. QALY: quality-adjusted life-year. €: Euros.

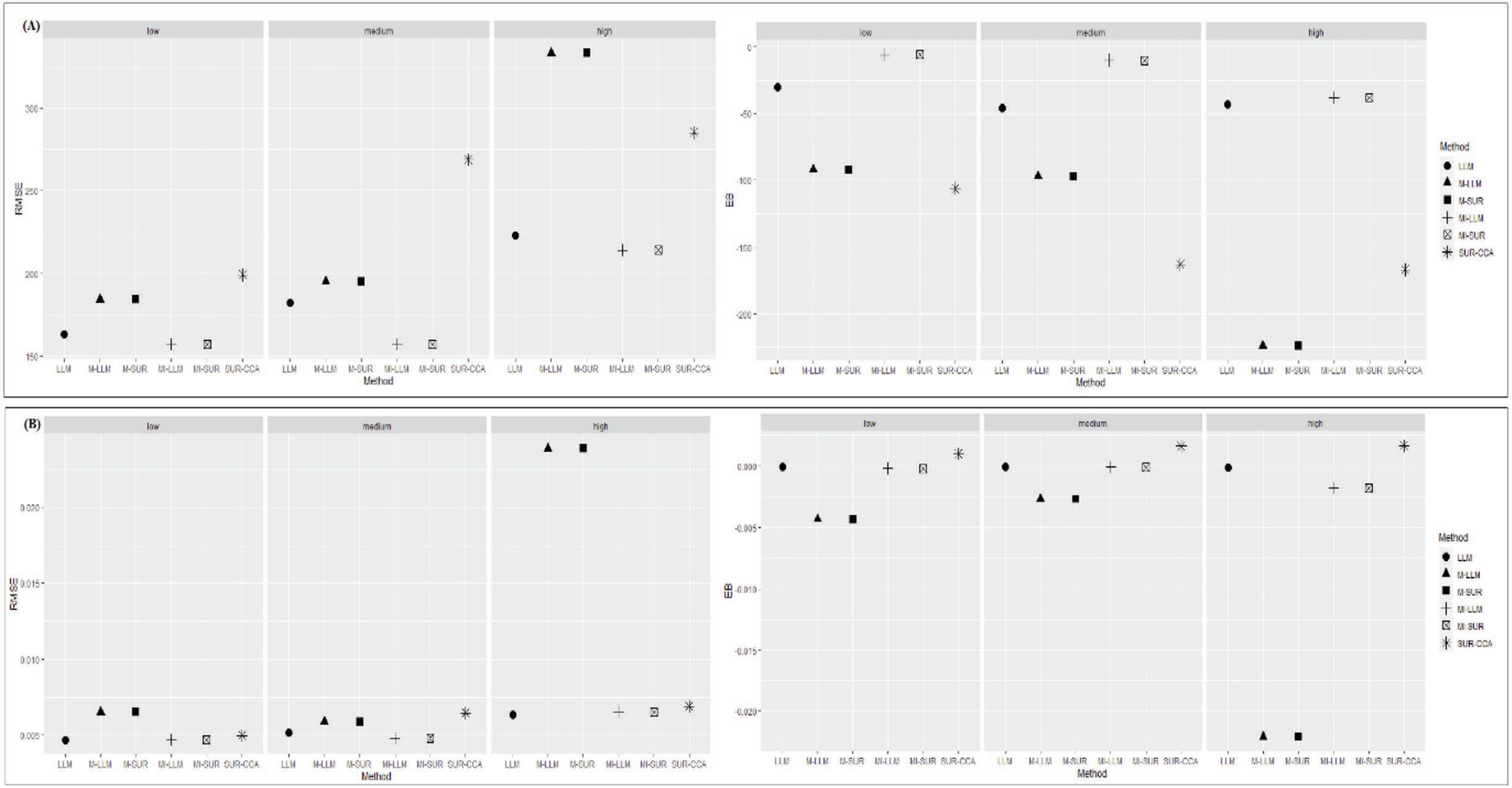


Figure 1 | Plots showing root-mean-square error (RMSE) and empirical bias (EB) associated with the six methodological strategies (Method) at different proportions of missing data (i.e., low: 10%, medium: 25%, and high: 50%) for costs (A) and QALY (B). QALY: quality-adjusted life-year. LLM: longitudinal mixed-model. M-LLM: Mean imputation combined with LLM. MI-LLM: Multiple imputation combined with LLM. SUR-CCA: Seemingly unrelated regressions - complete case analysis**.** M-SUR: mean imputation combined with SUR. MI-SUR: multiple imputation combined with SUR.

Empirical datasets

In empirical dataset 1, LLM, MI-LLM, and MI-SUR resulted in similar point estimates for QALYs, but not for costs. The 95% CIs around the cost and QALY differences were somewhat wider for LLM and MI-LLM compared to MI-SUR, but all 95% CIs showed that both differences were not statistically significant. Similar ICERs and probabilities of cost-effectiveness were found for LLM, MI-LLM, and MI-SUR. The mean imputation methods (M-LLM and M-SUR) resulted in narrower 95% CIs compared to LLM MI-LLM, and MI-SUR, but similar probabilities of cost-effectiveness. SUR-CCA resulted in very different point estimates, ICER, and probabilities of cost-effectiveness compared to all others methodological strategies (Table 3; Figure 2).

Table 3 | Cost-effectiveness results for different proportions of missing data in empirical datasets

Missing	Method	Δ Costs, € (95% CI)	Δ QALY (95% CI)	ICER, €/QALY	Probability of cost-effectiveness			
					€0/ QALY	€10,000/ QALY	€20,000/ QALY	€50,000/ QALY
Empirical dataset 1, N=169								
13%	**LLM**	-1048 (-2391; 296)	-0.002 (-0.072; 0.068)	441749	0.937	0.882	0.812	0.671
	M-LLM	-1073 (-2290; 143)	-0.001 (-0.069; 0.066)	762552	0.958	0.909	0.839	0.693
	MI-LLM	-1046 (-2391; 298)	-0.002 (-0.072; 0.068)	536379	0.936	0.882	0.813	0.674
	SUR-CCA, N=141	-332 (-1324; 660)	-0.005 (-0.029; 0.020)	70233	0.744	0.686	0.636	0.538
	M-SUR	-681 (-1569; 207)	-0.001 (-0.023; 0.020)	489833	0.934	0.900	0.861	0.755
	MI-SUR	-711 (-1674; 252)	-0.002 (-0.025; 0.022)	437230	0.926	0.891	0.851	0.746
Empirical dataset 2, N=143								
66%	**LLM**	-630 (-3773; 2513)	-0.018 (-0.103; 0.067)	35272	0.653	0.606	0.559	0.462
	M-LLM	-356 (-2089; 1376)	-0.030 (-0.115; 0.055)	11785	0.657	0.523	0.421	0.312
	MI-LLM	-816 (-3431; 1798)	-0.031 (-0.129; 0.067)	26242	0.730	0.638	0.546	0.398
	SUR-CCA, N=45	798 (-1299; 2894)	-0.017 (-0.090; 0.057)	-47329	0.228	0.204	0.203	0.233
	M-SUR	-309 (-1395; 777)	-0.018 (-0.058; 0.021)	17132	0.717	0.585	0.470	0.307
	MI-SUR	-763 (-2533; 1008)	-0.018 (-0.067; 0.030)	41151	0.800	0.729	0.646	0.458

LLM: longitudinal mixed-model. M-LLM: Mean imputation combined with LLM. MI-LLM: Multiple imputation combined with LLM. SUR-CCA: Seemingly unrelated regressions - complete case analysis**.** M-SUR: mean imputation combined with SUR. MI-SUR: multiple imputation combined with SUR. Δ: difference. CI: confidence interval. ICER: incremental cost-effectiveness ratio. QALY: quality-adjusted life-year. €: Euros.

In empirical dataset 2, as in empirical dataset 1, similar point estimates for QALY, but not for costs were found for LLM, MI-LLM, and MI-SUR. Results of SUR-CCA, M-SUR, and M-LLM were most different from those of LLM, MI-LL, and MI-SUR (Table 3). Figure 2 shows that the probabilities of cost-effectiveness of the LLM, M-LLM, MI-LLM, M-SUR, and MI-SUR somewhat differed, particularly at lower WTP thresholds.

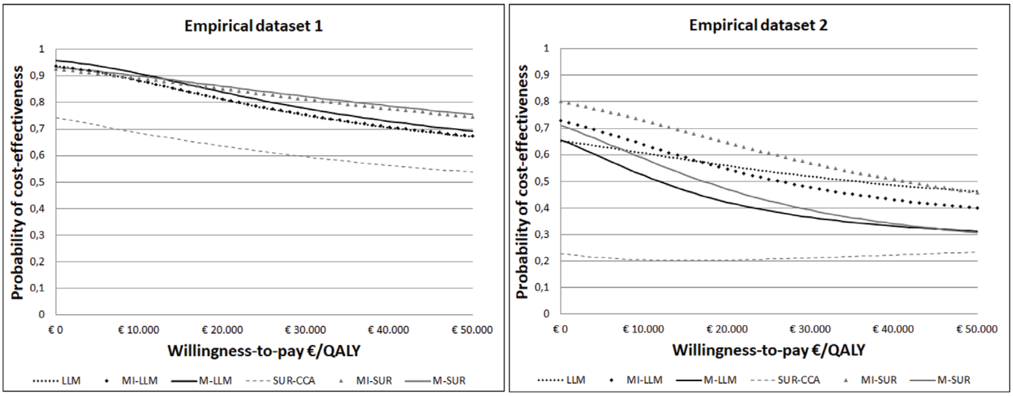


Figure 2 | Cost-effectiveness acceptability curves (CEACs) showing the probability of the intervention being cost-effective (x-axis) for different willingness-to-pay thresholds per unit of quality-adjusted life-year (QALY) gained (y-axis) in empirical datasets 1 and 2 with 9% and 53% of missing data in costs and QALY, respectively. LLM: longitudinal mixed-model. M-LLM: Mean imputation combined with LLM. MI-LLM: Multiple imputation combined with LLM. SUR-CCA: Seemingly unrelated regressions – complete case analysis**.** M-SUR: mean imputation combined with SUR. MI-SUR: multiple imputation combined with SUR. €: Euros.

Discussion

Main findings

Our findings suggest that MI prior to LLM does improve the method's performance for costs. However, at 50% missing data, all methods had a relatively high level of bias for costs. For QALY, LLM alone already has an acceptable level of performance. Furthermore, the performance of MI-LLM and MI-SUR was similar for costs as well as QALY. Finally, LLM, MI-LLM, and MI-SUR were found to perform considerably better than both mean imputation approaches (i.e., M-LLM, M-SUR) and SUR-CCA for costs and QALY. In empirical datasets, we found LLM, MI-LLM, and MI-SUR to result in similar point estimates for QALY, but not for costs, while the results of SUR-CCA, M-SUR and M-LLM were extensively different.

Interpretation of findings and comparison with the literature

Previous studies suggest that when using LLM, MI of missing values is not necessary to obtain unbiased effect estimates regardless of the missing data mechanism.[14,15] In line with these studies, we found MI-LLM and LLM to perform equally well for QALY.[14,15] For costs, however, we found MI prior to LLM to perform better than LLM alone. This difference in performance is likely because costs are typically associated with higher levels of skewness, kurtosis, within-subject variability, and between-subject variability compared with QALY. This phenomenon was also observed in our simulated as well as our empirical datasets. The LLM Stata command we used in our study assumes multivariate normality, whereas the PMM method used to impute data assumes that the distribution of missing values is the same as the observed one, which may result in more robustness to normality violations than maximum likelihood-based methods. According to Dong et al. (2013), this is so because violation of the multivariate normality assumption may cause convergence problems for maximum likelihood-based methods, while the posterior distribution in MI is approximated by a finite mixture of the normal distributions, enabling MI to capture non-normal features (e.g., skewness).[42] To the best of our knowledge, however, systematic comparisons of MI and maximum likelihood-based methods in terms of their sensitivity to the violation of the multivariate normality assumption are lacking.[42-44] Recently, Gabrio et al. (2022) assessed the applicability of LLM in one empirical trial-based economic evaluation showing similar point estimates for costs and QALY between LLM and MI-LLM.[31] In another study, Gabrio et al. (2021)[28] showed that a longitudinal model alone resulted in unbiased estimates under MAR in the context of a trial-based economic evaluation. Gabrio et al. (2021), however, considered Bayesian joint longitudinal models, rather than mixed effects models, and did not assess their methods' performance for costs and QALY separately.[28] In contrast to Gabrio et al. (2021) we did not consider any correlation between costs and QALY in our maximum likelihood-based models, whereas costs and effects are typically correlated in trial-based economic evaluations. In a post-hoc analysis we, therefore, assessed whether our maximum likelihood-based models would perform better when specifying them according to the suggestion of Faria et al. (2014).[5] That is, after rescaling costs to the same scale as utility values (i.e. 0-1), both outcomes were stacked on top of each other and simultaneously regressed upon the various covariates in the model using a three-level structure (i.e. subject, outcome, time)(see Supplementary material 5).[5] However, even though we did find that such a joint estimation of total costs and QALY slightly improved the models' bias, their coverage rates were found to be highly sensitive to an incorrect rescaling of costs, which makes the approach hard to apply in practice (Supplementary material 5). We also found MI-LLM to perform equally well as MI-SUR, while the former does not consider the possible correlation between costs and QALY and the latter does. This is in line with the results of Mutubuki et al. (2021) who found accounting for the correlation between costs and effects not to have a large impact on cost and QALY estimates in two empirical datasets, nor on the statistical uncertainty surrounding both outcomes.[45]

Strengths and limitations

To the best of our knowledge, our study is the first to assess whether it is necessary to perform MI before LLM to account for missing data in trial-based economic evaluations. Another strength of this study was the use of simulated data, which means that we know the 'true' outcomes and can, therefore, assess the statistical performance of the methods. In addition, we calculated the number of simulated datasets needed to draw valid conclusions and the simulated datasets resembled empirical data as closely as possible.[30]

This study also has some limitations. First, as previously discussed, LLM assumes multivariate normality, an assumption that is typically violated for costs. Future research should therefore assess whether Bayesian joint longitudinal models are better suited for analysing longitudinal trial-based economic evaluation data than the methods assessed in our study.[28] This might be because Bayesian models enable the joint estimation of costs and effects, while also allowing the use of different distributions for both outcomes (e.g. Gamma for costs and Beta for QALY).[28,46] Second, generating a strong MAR mechanism in simulated datasets with a longitudinal structure is relatively complex because a modification in one parameter impacts on all other parameters. To partially overcome this issue, slight baseline imbalances were introduced for age and gender. This, however, is in contrast with the theoretical advantage of RCTs, namely that the randomization of participants allows researchers the confidence that – on average – treatment groups are similar and that the only difference between both groups is the intervention to be assessed for its (cost-)effectiveness.[47] Nonetheless, every individual trial – by definition – exhibits some form of imbalance with respect to measured prognostic variables, especially smaller trials. Moreover, when generating the imbalances we made sure that they were small and in line with the slight baseline imbalances encountered in our empirical datasets.[47] Third, LLM can only deal with missing values at the aggregate level (i.e., total costs and utility values), which might make it less suitable when values are missing at the item level (e.g., number of GP visits, EQ-5D mobility dimension). Further research is therefore needed to assess whether the current results would hold when data are missing at the item-level. Fourth, four follow-up time points were simulated with equal time intervals (i.e., 3, 6, 9, and 12-months follow-up), whereas this is not necessarily the case in trial-based economic evaluations.[48] Further research is needed to assess the impact of varying time intervals on the performance of the methods assessed in this study. Fifth, in this study, we looked at MAR conditional on baseline values, whereas MAR can also be conditional on other values in the dataset. However, we do not expect our conclusion to change for MAR conditional at other values, because such other values can easily be included in an imputation model as well. Sixth, another limitation of the study is that results have only been assessed under MAR mechanisms, while MNAR is also possible. Since it is never possible to distinguish between MAR and MNAR from the data at hand, the robustness of deviations from the MAR assumption is ideally assessed in trial-based economic evaluations using sensitivity analyses. In doing so, the magnitude and direction of the departures from MAR are ideally defined based on external information (e.g. expert opinion).[5,7] Seventh, it is also relevant to mention that despite simulating skewed cost and effect data, we assumed that with the current sample size (i.e. 600 subjects), LMM estimates would be robust to non-normal distribution[49]. Amongst others, we did so because using bootstrapping to deal with skewed data made our simulations slower and time consuming. In addition, we wanted to ensure that the differences in performance could be attributed to the use of MI and/or LLM, and not to bootstrapping, and consensus does not currently exist as to how MI ought to be combined with bootstrapping.[50,51]

Implications for Practice and Research

MI-LLM was found to perform better than LLM in the simulated datasets and equally well as MI-SUR. Therefore, if researchers want to use LLM for analysing trial-based economic evaluation data they are advised to multiply impute missing values first. Next to the identified improved performance of LLM, MI also allows to include values in the imputation model that may be not relevant for the analysis model (i.e., auxiliary variables), allows to impute missing values at the item-level, and allows for a wider range of regression models to be used in the analysis (e.g., SUR).[23] From an efficiency standpoint, one might also opt for MI-SUR instead of MI-LLM, because it is computationally more efficient, while having similar EBs and RMSEs and CRs that are closer to the nominal value. Nevertheless, one should bear in mind, that SUR does not allow for the estimation of the average intervention effect over time, whereas LLM does. On the other hand, SUR account for the correlation between costs and effects, whereas LLM does not. Another important point is that variables included in the LLM model should have complete data at baseline otherwise, the model will exclude missing observations from the analysis. To facilitate researchers in using either one of these methods, software codes are included to this manuscript as Supplementary materials 3, and 5. Moreover, as the methods' performance was found to be considerably less with 50% missing data, researchers are advised to extensively assess the robustness of their results to the methods applied for handling missing data, particularly with high percentages of missing data.

As previously discussed, it should be noted that the presence of item-level missingness highly depends on the measures used to collect the data. For example, it is rarely the case to observe item-level missingness in short questionnaires, such as the EQ-5D, which almost always exhibits monotone patterns.[29] On the contrary, resource use questionnaires often present item-level missingness in trial-based economic evaluations, which makes non-monotone pattern likely to be more relevant for costs. In the current study, we only simulated monotone patterns, hence further investigation is needed to assess whether the current results hold when data are missing at the item level and/or when time intervals vary over time.

Further research is also needed to assess whether Bayesian joint longitudinal models are preferred over the frequentist models evaluated in the current study. Among others, the relevance of exploring the performance of Bayesian models are that they can encode missingness assumptions via prior distributions (e.g., MNAR) and quantify uncertainty without the need to implement bootstrapping methods.[52,53]

Conclusions

Our findings suggest that if researchers want to use LLM for analysing trial-based economic evaluation data, it is advisable to multiply impute missing values first. One might also opt for a combination of MI and SUR. MI-SUR is computationally more efficient than MI-LLM, but it does not allow for the estimation of average intervention effects over time while having the advantage of accounting for the correlation between costs and effects. Our findings also underscore the importance of extensively assessing the robustness of results to the methods applied for handling missing data, particularly with high percentages of missing data. Further research should assess the relative performance of MI-LLM and MI-SUR versus Bayesian joint longitudinal models, under MAR and MNAR assumptions and assess whether the current results hold when data are missing at the item-level and/or when time intervals vary over time.

References

1. WHO: 2015 Global Survey on Health Technology Assessment by National Authorities, https://www.who.int/health-technology-assessment/examples/en/, (2015)
2. Drummond, M.F., Sculpher, M.J., Torranc, G.W.: Methods for the economic evaluation of health care programmes. Oxford University Press, Oxford (2005)
3. Petrou, S., Gray, A.: Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. The BMJ. 342, (2011). https://doi.org/10.1136/bmj.d1548
4. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons (2004)
5. Faria, R., Gomes, M., Epstein, D., White, I.R.: A Guide to Handling Missing Data in Cost-Effectiveness Analysis Conducted Within Randomised Controlled Trials. Pharmacoeconomics. 32, 1157 (2014). https://doi.org/10.1007/s40273-014-0193-3
6. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. John Wiley & Sons, Inc., New York, NY, USA (2014)
7. Gabrio, A., Mason, A.J., Baio, G.: Handling Missing Data in Within-Trial Cost-Effectiveness Analysis: A Review with Future Recommendations. PharmacoEconomics - Open. 1, 79–97 (2017). https://doi.org/10.1007/s41669-017-0015-6
8. Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J.G., Sinha, D., Parzen, M., Lipshultz, S.: Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: An application to AIDS data. J. R. Stat. Soc. Ser. A Stat. Soc. 172, 3–20 (2009). https://doi.org/10.1111/j.1467-985X.2008.00564.x
9. Noble, S.M., Hollingworth, W., Tilling, K.: Missing data in trial-based cost-effectiveness analysis: the current state of play. Health Econ. 21, 187–200 (2012). https://doi.org/10.1002/hec.1693
10. Díaz-Ordaz, K., Kenward, M.G., Cohen, A., Coleman, C.L., Eldridge, S.: Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. Clin. Trials Lond. Engl. 11, 590–600 (2014). https://doi.org/10.1177/1740774514537136
11. Zhang, Z.: Missing data imputation: focusing on single imputation. Ann. Transl. Med. 4, (2016). https://doi.org/10.3978/j.issn.2305-5839.2015.12.38
12. Gomes, M., Grieve, R., Nixon, R., Edmunds, W.J.: Statistical Methods for Cost-Effectiveness Analyses That Use Data from Cluster Randomized Trials: A Systematic Review and Checklist for Critical Appraisal. Med. Decis. Making. 32, 209–220 (2012). https://doi.org/10.1177/0272989X11407341
13. Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., Carpenter, J.R.: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 338, b2393 (2009). https://doi.org/10.1136/bmj.b2393
14. Twisk, J., de Boer, M., de Vente, W., Heymans, M.: Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. J. Clin. Epidemiol. 66, 1022–1028 (2013). https://doi.org/10.1016/j.jclinepi.2013.03.017
15. Peters, S.A.E., Bots, M.L., Ruijter, H.M. den, Palmer, M.K., Grobbee, D.E., Crouse, J.R., O'Leary, D.H., Evans, G.W., Raichlen, J.S., Moons, K.G.M., Koffijberg, H.: Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models. J. Clin. Epidemiol. 65, 686–695 (2012). https://doi.org/10.1016/j.jclinepi.2011.11.012
16. Willan, A.R., Briggs, A.H., Hoch, J.S.: Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. Health Econ. 13, 461–475 (2004). https://doi.org/10.1002/hec.843
17. Sullivan, T.R., White, I.R., Salter, A.B., Ryan, P., Lee, K.J.: Should multiple imputation be the method of choice for handling missing data in randomized trials? Stat. Methods Med. Res. 27, 2610–2626 (2018). https://doi.org/10.1177/0962280216683570

18. Oosterhuis, T., Ostelo, R.W., van Dongen, J.M., Peul, W.C., de Boer, M.R., Bosmans, J.E., Vleggeert-Lankamp, C.L., Arts, M.P., van Tulder, M.W.: Early rehabilitation after lumbar disc surgery is not effective or cost-effective compared to no referral: a randomised trial and economic evaluation. J. Physiother. 63, 144–153 (2017). https://doi.org/10.1016/j.jphys.2017.05.016

19. Bosmans, J.E., van Schaik, D.J.F., Heymans, M.W., van Marwijk, H.W.J., van Hout, H.P.J., de Bruijne, M.C.: Cost-effectiveness of interpersonal psychotherapy for elderly primary care patients with major depression. Int. J. Technol. Assess. Health Care. 23, 480–487 (2007). https://doi.org/10.1017/S0266462307070572

20. Goldfeld, K.: simstudy: Simulation of Study Data, https://CRAN.R-project.org/package=simstudy, (2019)

21. Wit, M., Rondags, S.M.P.A., Tulder, M.W., Snoek, F.J., Bosmans, J.E.: Cost-effectiveness of the psycho-educational blended (group and online) intervention HypoAware compared with usual care for people with Type 1 and insulin-treated Type 2 diabetes with problematic hypoglycaemia: analyses of a cluster-randomized controlled trial. Diabet. Med. 35, 214–222 (2018). https://doi.org/10.1111/dme.13548

22. Pols, A.D., van Dijk, S.E., Bosmans, J.E., Hoekstra, T., van Marwijk, H.W.J., van Tulder, M.W., Adriaanse, M.C.: Effectiveness of a stepped-care intervention to prevent major depression in patients with type 2 diabetes mellitus and/or coronary heart disease and subthreshold depression: A pragmatic cluster randomized controlled trial. PLoS ONE. 12, (2017). https://doi.org/10.1371/journal.pone.0181023

23. Whitehead, S.J., Ali, S.: Health outcomes in economic evaluation: the QALY and utilities. Br. Med. Bull. 96, 5–21 (2010). https://doi.org/10.1093/bmb/ldq033

24. Schouten, R.M., Lugtig, P., Vink, G.: Generating missing values for simulation purposes: a multivariate amputation procedure. J. Stat. Comput. Simul. 88, 2909–2930 (2018). https://doi.org/10.1080/00949655.2018.1491577

25. Apeldoorn, A.T., Bosmans, J.E., Ostelo, R.W., de Vet, H.C.W., van Tulder, M.W.: Cost-effectiveness of a classification-based system for sub-acute and chronic low back pain. Eur. Spine J. 21, 1290–1300 (2012). https://doi.org/10.1007/s00586-011-2144-4

26. El Alili, M., van Dongen, J.M., Goldfeld, K.S., Heymans, M.W., van Tulder, M.W., Bosmans, J.E.: Taking the Analysis of Trial-Based Economic Evaluations to the Next Level: The Importance of Accounting for Clustering. PharmacoEconomics. (2020). https://doi.org/10.1007/s40273-020-00946-y

27. Flynn, T., Peters, T.: Conceptual issues in the analysis of cost data within cluster randomized trials. J. Health Serv. Res. Policy. 10, 97–102 (2005). https://doi.org/10.1258/1355819053559065

28. Gabrio, A., Hunter, R., Mason, A.J., Baio, G.: Joint Longitudinal Models for Dealing With Missing at Random Data in Trial-Based Economic Evaluations. Value Health. 24, 699–706 (2021). https://doi.org/10.1016/j.jval.2020.11.018

29. Simons, C.L., Rivero-Arias, O., Yu, L.-M., Simon, J.: Multiple imputation to deal with missing EQ-5D-3L data: Should we impute individual domains or the actual index? Qual. Life Res. 24, 805–815 (2015). https://doi.org/10.1007/s11136-014-0837-y

30. Morris, T.P., White, I.R., Crowther, M.J.: Using simulation studies to evaluate statistical methods. Stat. Med. 38, 2074–2102 (2019). https://doi.org/10.1002/sim.8086

31. Gabrio, A., Plumpton, C., Banerjee, S., Leurent, B.: Linear mixed models to handle missing at random data in trial-based economic evaluations. Health Econ. 31, 1276–1287 (2022). https://doi.org/10.1002/hec.4510

32. Twisk, J.W.R.: Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. Cambridge University Press (2013)

33. White, I.R., Royston, P., Wood, A.M.: Multiple imputation using chained equations: Issues and guidance for practice. Stat. Med. 30, 377–399 (2011). https://doi.org/10.1002/sim.4067

34. Morris, T.P., White, I.R., Royston, P.: Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med. Res. Methodol. 14, 75 (2014). https://doi.org/10.1186/1471-2288-14-75

35. Zellner, A.: Estimators for Seemingly Unrelated Regression Equations: Some Exact Finite Sample Results. J. Am. Stat. Assoc. 58, 977–992 (1963). https://doi.org/10.1080/01621459.1963.10480681

36. Collins, L.M., Schafer, J.L., Kam, C.-M.: A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol. Methods. 6, 330–351 (2001). https://doi.org/10.1037/1082-989X.6.4.330

37. Pederson, S.P., Forster, R.A., Booth, T.E.: Confidence Interval Procedures for Monte Carlo Transport Simulations. Nucl. Sci. Eng. 127, 54–77 (1997). https://doi.org/10.13182/NSE97-A1921

38. Yucel, R.M., Demirtas, H.: Impact of non-normal random effects on inference by multiple imputation: A simulation assessment. Comput. Stat. Data Anal. 54, 790–801 (2010). https://doi.org/10.1016/j.csda.2009.01.016

39. Fenwick, E., O'Brien, B.J., Briggs, A.: Cost-effectiveness acceptability curves – facts, fallacies and frequently asked questions. Health Econ. 13, 405–415 (2004). https://doi.org/10.1002/hec.903

40. Löthgren, M., Zethraeus, N.: Definition, interpretation and calculation of cost-effectiveness acceptability curves. Health Econ. 9, 623–630 (2000)

41. Ministerie van Volksgezondheid, W. en S.: Evaluatie van eHealth-technologie - Publicatie - Zorginstituut Nederland. Ministerie van Volksgezondheid, Welzijn en Sport, https://www.zorginstituutnederland.nl/publicaties/publicatie/2017/05/17/evaluatie-van-ehealth-technologie

42. Dong, Y., Peng, C.-Y.J.: Principled missing data methods for researchers. SpringerPlus. 2, 222 (2013). https://doi.org/10.1186/2193-1801-2-222

43. Schafer, J.L.: Multiple imputation: a primer. Stat. Methods Med. Res. 8, 3–15 (1999). https://doi.org/10.1177/096228029900800102

44. Knorr-Held, L.: Analysis of Incomplete Multivariate Data. J. L. Schafer, Chapman & Hall, London, 1997. No. of pages: xiv+430. Price: £39.95. ISBN 0-412-04061-1. Stat. Med. 19, 1006–1008 (2000). https://doi.org/10.1002/(SICI)1097-0258(20000415)19:7<1006::AID-SIM384>3.0.CO;2-T

45. Mutubuki, E.N., El Alili, M., Bosmans, J.E., Oosterhuis, T., J Snoek, F., Ostelo, R.W.J.G., van Tulder, M.W., van Dongen, J.M.: The statistical approach in trial-based economic evaluations matters: get your statistics together! BMC Health Serv. Res. 21, 475 (2021). https://doi.org/10.1186/s12913-021-06513-1

46. Manca, A., Lambert, P.C., Sculpher, M., Rice, N.: Cost-effectiveness analysis using data from multinational trials: the use of bivariate hierarchical modeling. Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak. 27, 471–490 (2007). https://doi.org/10.1177/0272989X07302132

47. Ciolino, J.D., Palac, H.L., Yang, A., Vaca, M., Belli, H.M.: Ideal vs. real: a systematic review on handling covariates in randomized controlled trials. BMC Med. Res. Methodol. 19, 136 (2019). https://doi.org/10.1186/s12874-019-0787-8

48. Huque, M.H., Carlin, J.B., Simpson, J.A., Lee, K.J.: A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med. Res. Methodol. 18, 168 (2018). https://doi.org/10.1186/s12874-018-0615-6

49. Nixon, R.M., Wonderling, D., Grieve, R.D.: Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. Health Econ. 19, 316–333 (2010). https://doi.org/10.1002/hec.1477

50. Schomaker, M., Heumann, C.: Bootstrap inference when using multiple imputation. Stat. Med. 37, 2252–2266 (2018). https://doi.org/10.1002/sim.7654

51. Brand, J., van Buuren, S., le Cessie, S., van den Hout, W.: Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. Stat. Med. 38, 210–220 (2019). https://doi.org/10.1002/sim.7956

52. Gabrio, A., Daniels, M.J., Baio, G.: A Bayesian parametric approach to handle missing longitudinal outcome data in trial-based health economic evaluations. J. R. Stat. Soc. Ser. A Stat. Soc. 183, 607–629 (2020). https://doi.org/10.1111/rssa.12522

53. Mason, A.J., Gomes, M., Carpenter, J., Grieve, R.: Flexible Bayesian longitudinal models for cost-effectiveness analyses with informative missing data. Health Econ. 30, 3138–3158 (2021). https://doi.org/10.1002/hec.4408

54. Coretti, S., Ruggeri, M., McNamee, P.: The minimum clinically important difference for EQ-5D index: a critical review. Expert Rev. Pharmacoecon. Outcomes Res. 14, 221–233 (2014). https://doi.org/10.1586/14737167.2014.894462

55. Walters, S.J., Brazier, J.E.: Comparison of the Minimally Important Difference for Two Health State Utility Measures: EQ-5D and SF-6D. Qual. Life Res. 14, 1523–1532 (2005)

SUPPLEMENTARY MATERIAL 1

Data generating code in R

```
# _Simulation of longitudinal data for utilities and costs including covariates at baseline_

## COMPLETE DATA
library(simstudy) # required to generate correlated longitudinal data
library(foreign) # required to save data in .dta

for (i in 1:2000){
 set.seed(i)
 ##Generate baseline characteristics
 def <- defData(varname = “variance”, formula = 0.005, dist = “normal”)
 def <- defData(def, varname = “gammaDisC”, formula = 0.1, dist = “nonrandom”)
 def <- defData(def, varname = “nMeasurement”, formula = 1, dist = “nonrandom”, id = “id”)
 def <- defData(def, varname = “trt”, formula = 0.52, dist = “binary”, id=”id”)
 data <- genData(600, def)

 defb <- defDataAdd(varname = “age”, formula = “40 + (5*trt)”, variance = 140, dist=”normal”)
 defb <- defDataAdd(defb, varname = “gender”, formula = “0.52 + (0.4*trt)”, dist = “binary”)
 data <- addCorFlex(data, defb, rho = 0.0001, corstr = “cs”)
 data$age <- ifelse(data$age <= 18, 18, data$age)
 data$age <- ifelse(data$age >= 99, 99, data$age)

 defb <- defDataAdd(varname = “uT0”, formula = “0.2 + (0.0045 * age) + (0.025 * gender)”, variance = 0.002, dist =
“gamma”, link = “logit”)
 defb <- defDataAdd(defb, varname = “cT0”, formula = “50 + (15*age) + (100*gender)”, variance = 0.15, dist = “gamma”,
link = “logit”)
 data <- addCorFlex(data, defb, rho = -0.5)

 ##Generate follow-up utilities and costs
 defc <- defDataAdd(varname = “uT1”, formula = “uT0 + (0.04* trt) + (0.002 * age) + (0.01 * gender)”, variance = 0.002,
dist = “gamma”, link = “logit”)
 defc <- defDataAdd(defc, varname = “cT1”, formula = “cT0 + (62.5*trt) + (1*age) + (10*gender)”, variance = 0.15, dist
= “gamma”, link = “logit”)
 data <- addCorFlex(data, defc, rho = -0.5)
 data <- genCluster(data, “id”, “nMeasurement”, “idcluster”)

 ##Correlation between time points for utilities
 Q <- matrix(c(1.0,0.9,0.9,0.9,1.0,0.9,0.9,0.9,1.0), nrow = 3)
 data <- addCorGen(dtOld = data, idvar = “id”, nvars = 3, corMatrix = Q, dist = “normal”, param1 = “uT1”, param2 =
“variance”, cnames = “uT2, uT3, uT4”)

 ##Correlation between time points for costs
 C <- matrix(c(1.0,0.7,0.7,0.7,1.0,0.7,0.7,0.7,1.0), nrow = 3)
 data <- addCorGen(dtOld = data, idvar = “id”, nvars = 3, corMatrix = C, dist = “gamma”, param1 = “cT1”, param2 =
“gammaDisC”, cnames = “cT2, cT3, cT4”)
 write.dta(data, file = paste0(“C:/completedata/dataset”,i,”.dta”))
}
```

Note: the correlations and the coefficients set corresponded approximately to the values generated due to relative complexity of the simulated longitudinal data.

SUPPLEMENTARY MATERIAL 2

To generate follow-up missing values, a missing data indicator m_{ij} was generated for every subject i (i = 1, …, N=600) at time point j (j = 1, …, 4) using a binomial distribution and a logit link function to model the linear dependence between the probability of missing data and its predictive variables at baseline (i.e., age, gender, cT0, uT0, trt).

$$m_{ij} \sim \ Binomial(\pi_{mij}),$$

$$logit(\pi_{mij}) = \ \beta_0 + \ \beta_1 age_i + \ \beta_2 gender_i + \ \beta_3 uT0_i + \ \beta_4 cT0_i + \ \beta_5 trt_i + \ \varepsilon_i, \qquad (1)$$

where π_{mij} is the probability of a subject i having missing data at time point j. The intercept (β_0) and coefficients (β_1 ,β_2,β_3,β_4,β_5) of covariates were tweaked to generate a plausible MAR assumption and datasets with 10%, 25%, and 50% missings, and ε_i is the error term (Supplementary material 2)[28]. Then, at the respective time point j, all complete follow-up cost and utility values were replaced by missing values according to m_{ij}. This process was repeated for all consecutive time points amongst the subset of individuals still in the study.[28]

Supplementary Table 1 | Parameter values used for generating missing data

Mechanism, missing %	Parameter values
MAR, 10%	
T1	$if\ trt = 1, \beta_1 = 0.110; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_1 = 0.110; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.0001; \beta_5 = 1$
T2	$if\ trt = 1, \beta_0 = -0.1; \beta_1 = 0.100; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_0 = -0.1; \beta_1 = 0.100; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.00027; \beta_5 = 1$
T3	$if\ trt = 1, \beta_0 = -0.2; \beta_1 = 0.90; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_0 = -0.2; \beta_1 = 0.80; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.0001; \beta\mathrm{v} = 1$
T4	$if\ trt = 1, \mathrm{v} = -1.9; \beta_1 = 0.100; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_0 = -1.9; \beta_1 = 0.50; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.0001; \beta_5 = 1$
MAR 25%	
T1	$if\ trt = 1, \beta_1 = 0.110; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_1 = 0.110; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.0001; \beta_5 = 1$
T2	$if\ trt = 1, \beta_0 = -0.1; \beta_1 = 0.100; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_0 = -0.1; \beta_1 = 0.100; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.00027; \beta_5 = 1$
T3	$if\ trt = 1, \beta_0 = -0.1; \beta_1 = 0.100; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_0 = -0.1; \beta_1 = 0.033; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.0001; \beta_5 = 1$
T4	$if\ trt = 1, \beta_1 = 0.300; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_1 = -0.300; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.0001; \beta_5 = 1$
MAR 50%	
T1	$if\ trt = 1, \beta_1 = 0.110; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_1 = 0.110; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.0001; \beta\mathrm{v} = 1$
T2	$if\ trt = 1, \beta_0 = -0.1; \beta_1 = 0.100; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_0 = -0.1; \beta_1 = 0.100; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.00027; \beta_5 = 1$
T3	$if\ trt = 1, \beta_0 = -0.1; \beta_1 = 0.100; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_0 = -0.1; \beta_1 = 0.033; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.0001; \beta_5 = 1$
T4	$if\ trt = 1, \beta_1 = 0.300; \beta_2 = 1; \beta_3 = 0.65; \beta_4 = 0.00027; \beta_5 = 1$
	$if\ trt = 0, \beta_1 = -0.300; \beta_2 = -1; \beta_3 = -0.65; \beta_4 = -0.0001; \beta_5 = 1$

Missing data generating code in R

```
library(foreign) # required to save data in .dta
library(descr)
library(dplyr)
## Import 2000 simulated datasets
datasets <- paste0(“C:/completedata/dataset”,1:2000,”.dta”)
data <- mclapply(datasets, read.dta)

MISSING10 <- function(ds){
 set.seed(2000)
 ds$totalcost <- sum(ds$cT1 + ds$cT2 + ds$cT3 + ds$cT4)
 ds$mediantotalcost <- median(ds$totalcost)
 ds$QALYs <- sum(ds$uT1 + ds$uT2 + ds$uT3 + ds$uT4)/4
 ds$medianQALYs <- median(ds$QALYs)
 ds$z1 <- ifelse(ds$trt == 1, 0.110*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
 ds$z1 <- ifelse(ds$trt == 0, 0.110*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00010*ds$cT0 + 1*ds$trt, ds$z1)
 ds$pr1a = exp(-ds$z1)
 ds$pr1b = 1 + ds$pr1a
 ds$pr1c = 1 / ds$pr1b
 ds$y1 = rbinom(600,1,ds$pr1c)
 ds$uT1<-replace(ds$uT1, ds$y1==0,NA)
 ds$cT1<-replace(ds$cT1, ds$y1==0,NA)
 ds$M <- as.integer(complete.cases(ds$cT1))
 descr(ds$M)
 ds$uT2 <-replace(ds$uT2, ds$y1==0,NA)
 ds$cT2 <-replace(ds$cT2, ds$y1==0,NA)
 ds$z2 <- ifelse(ds$trt == 1, -0.1 + 0.100*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
 ds$z2 <- ifelse(ds$trt == 0, -0.1 + 0.100*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00027*ds$cT0 + 1*ds$trt, ds$z2)
 ds$pr2a = exp(-ds$z2)
 ds$pr2b = 1 + ds$pr2a
 ds$pr2c = 1 / ds$pr2b
 ds$y2 = rbinom(600,1,ds$pr2c)
 ds$uT2<-replace(ds$uT2, ds$y2==0,NA)
 ds$cT2<-replace(ds$cT2, ds$y2==0,NA)
 ds$M2 <- as.integer(complete.cases(ds$cT2))
 descr(ds$M2)
 ds$uT3<-replace(ds$uT3, ds$y1==0,NA)
 ds$cT3<-replace(ds$cT3, ds$y1==0,NA)
 ds$uT3<-replace(ds$uT3, ds$y2==0,NA)
 ds$cT3<-replace(ds$cT3, ds$y2==0,NA)
 ds$z3 <- ifelse(ds$trt == 1, -0.2 + 0.90*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
 ds$z3 <- ifelse(ds$trt == 0, -0.2 + 0.80*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00010*ds$cT0 + 1*ds$trt, ds$z3)
 ds$pr3a = exp(-ds$z3)
 ds$pr3b = 1 + ds$pr3a
 ds$pr3c = 1 / ds$pr3b
 ds$y3 = rbinom(600,1,ds$pr3c)
 ds$uT3<-replace(ds$uT3, ds$y3==0,NA)
 ds$cT3<-replace(ds$cT3, ds$y3==0,NA)
 ds$M3 <- as.integer(complete.cases(ds$cT3))
 descr(ds$M3)
 ds$uT4<-replace(ds$uT4, ds$y1==0,NA)
 ds$cT4<-replace(ds$cT4, ds$y1==0,NA)
 ds$uT4<-replace(ds$uT4, ds$y2==0,NA)
 ds$cT4<-replace(ds$cT4, ds$y2==0,NA)
 ds$uT4<-replace(ds$uT4, ds$y3==0,NA)
 ds$cT4<-replace(ds$cT4, ds$y3==0,NA)
 ds$z4 <- ifelse(ds$trt == 1, -1.9 + 0.100*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
 ds$z4 <- ifelse(ds$trt == 0, -1.9 + 0.050*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00010*ds$cT0 + 1*ds$trt, ds$z4)
 ds$pr4a = exp(-ds$z4)
 ds$pr4b = 1 + ds$pr4a
 ds$pr4c = 1/ ds$pr4b
 ds$y4 = rbinom(600,1,ds$pr4c)
 ds$uT4<-replace(ds$uT4, ds$trt==1 & ds$y4==0 & ds$uT4 >= ds$mediantotalcost,NA)
 ds$cT4<-replace(ds$cT4, ds$trt==0 & ds$y4==0 & ds$cT4 <= ds$medianQALYs,NA)
 descr(ds$y4)
 ds$M <- as.integer(complete.cases(ds))
 descr(ds$M)
 ds <- subset(ds,select =c(“id”,”trt”,”age”,”gender”,”uT0”,”cT0”,”uT1”,”cT1”,”uT2”,”cT2”,”uT3”,”cT3”,”uT4”,”cT4”))
 }
test <- lapply(data,MISSING10)
for (i in 1:2000) { write.dta(test[[i]], file = paste0(“C:/MISSING10/dataset”,i,”.dta”))}

MISSING25 <- function(ds){
 set.seed(2001)
 ds$totalcost <- sum(ds$cT1 + ds$cT2 + ds$cT3 + ds$cT4)
 ds$mediantotalcost <- median(ds$totalcost)
 ds$QALYs <- sum(ds$uT1 + ds$uT2 + ds$uT3 + ds$uT4)/4
 ds$medianQALYs <- median(ds$QALYs)
 ds$z1 <- ifelse(ds$trt == 1, 0.110*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
 ds$z1 <- ifelse(ds$trt == 0, 0.110*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00010*ds$cT0 + 1*ds$trt, ds$z1)
 ds$pr1a = exp(-ds$z1)
 ds$pr1b = 1 + ds$pr1a
 ds$pr1c = 1 / ds$pr1b
 ds$y1 = rbinom(600,1,ds$pr1c)
 ds$uT1<-replace(ds$uT1, ds$y1==0,NA)
 ds$cT1<-replace(ds$cT1, ds$y1==0,NA)
 ds$M <- as.integer(complete.cases(ds$cT1))
 descr(ds$M)
 ds$uT2 <-replace(ds$uT2, ds$y1==0,NA)
 ds$cT2 <-replace(ds$cT2, ds$y1==0,NA)
 ds$z2 <- ifelse(ds$trt == 1, -0.1 + 0.100*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
 ds$z2 <- ifelse(ds$trt == 0, -0.1 + 0.100*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00027*ds$cT0 + 1*ds$trt, ds$z2)
 ds$pr2a = exp(-ds$z2)
 ds$pr2b = 1 + ds$pr2a
 ds$pr2c = 1 / ds$pr2b
 ds$y2 = rbinom(600,1,ds$pr2c)
 ds$uT2<-replace(ds$uT2, ds$y2==0,NA)
 ds$cT2<-replace(ds$cT2, ds$y2==0,NA)
 ds$M2 <- as.integer(complete.cases(ds$cT2))
 descr(ds$M2)
 ds$uT3<-replace(ds$uT3, ds$y1==0,NA)
 ds$cT3<-replace(ds$cT3, ds$y1==0,NA)
 ds$uT3<-replace(ds$uT3, ds$y2==0,NA)
 ds$cT3<-replace(ds$cT3, ds$y2==0,NA)
 ds$z3 <- ifelse(ds$trt == 1, -0.1 + 0.100*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
 ds$z3 <- ifelse(ds$trt == 0, -0.1 + 0.033*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00010*ds$cT0 + 1*ds$trt, ds$z3)
 ds$pr3a = exp(-ds$z3)
 ds$pr3b = 1 + ds$pr3a
 ds$pr3c = 1 / ds$pr3b
 ds$y3 = rbinom(600,1,ds$pr3c)
 ds$uT3<-replace(ds$uT3, ds$y3==0,NA)
 ds$cT3<-replace(ds$cT3, ds$y3==0,NA)
 ds$M3 <- as.integer(complete.cases(ds$cT3))
 descr(ds$M3)
 ds$uT4<-replace(ds$uT4, ds$y1==0,NA)
 ds$cT4<-replace(ds$cT4, ds$y1==0,NA)
 ds$uT4<-replace(ds$uT4, ds$y2==0,NA)
 ds$cT4<-replace(ds$cT4, ds$y2==0,NA)
 ds$uT4<-replace(ds$uT4, ds$y3==0,NA)
 ds$cT4<-replace(ds$cT4, ds$y3==0,NA)
 ds$z4 <- ifelse(ds$trt == 1, + 0.300*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
 ds$z4 <- ifelse(ds$trt == 0, - 0.300*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00010*ds$cT0 + 1*ds$trt, ds$z4)
 ds$pr4a = exp(-ds$z4)
 ds$pr4b = 1 + ds$pr4a
 ds$pr4c = 1/ ds$pr4b
 ds$y4 = rbinom(600,1,ds$pr4c)
 ds$uT4<-replace(ds$uT4, ds$trt==1 & ds$y4==0 & ds$uT4 >= ds$mediantotalcost,NA)
 ds$cT4<-replace(ds$cT4, ds$trt==0 & ds$y4==0 & ds$cT4 <= ds$medianQALYs,NA)
 descr(ds$y4)
 ds$M4 <- as.integer(complete.cases(ds$cT4))
 descr(ds$M4)
 ds$M <- as.integer(complete.cases(ds))
 descr(ds$M)
 ds <- as.data.frame(ds)
 ds <- subset(ds,select =c(“id”,”trt”,”age”,”gender”,”uT0”,”cT0”,”uT1”,”cT1”,”uT2”,”cT2”,”uT3”,”cT3”,”uT4”,”cT4”))
 }
test <- lapply(data,MISSING25)
for (i in 1:2000) {write.dta(test[[i]], file = paste0(“C:/MISSING25/dataset”,i,”.dta”))}

MISSING50 <- function(ds){
 set.seed(2012)
 ds$totalcost <- sum(ds$cT1 + ds$cT2 + ds$cT3 + ds$cT4)
 ds$mediantotalcost <- median(ds$totalcost)
 ds$QALYs <- sum(ds$uT1 + ds$uT2 + ds$uT3 + ds$uT4)/4
 ds$medianQALYs <- median(ds$QALYs)
 ds$z1 <- ifelse(ds$trt == 1, -0.1 - 0.0715*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
 ds$z1 <- ifelse(ds$trt == 0, -0.1 + 0.128*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00010*ds$cT0 + 1*ds$trt, ds$z1)
 ds$pr1a = exp(-ds$z1)
```

```
ds$pr1b = 1 + ds$pr1a
ds$pr1c = 1 / ds$pr1b
ds$y1 = rbinom(600,1,ds$pr1c)
ds$uT1<-replace(ds$uT1, ds$y1==0,NA)
ds$cT1<-replace(ds$cT1, ds$y1==0,NA)
ds$M <- as.integer(complete.cases(ds$cT1))
descr(ds$M)
ds$uT2 <-replace(ds$uT2, ds$y1==0,NA)
ds$cT2 <-replace(ds$cT2, ds$y1==0,NA)
ds$z2 <- ifelse(ds$trt == 1, -0.1 + 0.105*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
ds$z2 <- ifelse(ds$trt == 0, -0.1 + 0.128*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00027*ds$cT0 + 1*ds$trt, ds$z2)
ds$pr2a = exp(-ds$z2)
ds$pr2b = 1 + ds$pr2a
ds$pr2c = 1 / ds$pr2b
ds$y2 = rbinom(600,1,ds$pr2c)
ds$uT2<-replace(ds$uT2, ds$y2==0,NA)
ds$cT2<-replace(ds$cT2, ds$y2==0,NA)
ds$M2 <- as.integer(complete.cases(ds$cT2))
descr(ds$M2)
ds$uT3<-replace(ds$uT3, ds$y1==0,NA)
ds$cT3<-replace(ds$cT3, ds$y1==0,NA)
ds$uT3<-replace(ds$uT3, ds$y2==0,NA)
ds$cT3<-replace(ds$cT3, ds$y2==0,NA)
ds$z3 <- ifelse(ds$trt == 1, -0.1 + 0.110*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
ds$z3 <- ifelse(ds$trt == 0, -0.1 + 0.130*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00010*ds$cT0 + 1*ds$trt, ds$z3)
ds$pr3a = exp(-ds$z3)
ds$pr3b = 1 + ds$pr3a
ds$pr3c = 1 / ds$pr3b
ds$y3 = rbinom(600,1,ds$pr3c)
ds$uT3<-replace(ds$uT3, ds$y3==0,NA)
ds$cT3<-replace(ds$cT3, ds$y3==0,NA)
ds$M3 <- as.integer(complete.cases(ds$cT3))
descr(ds$M3)
ds$uT4<-replace(ds$uT4, ds$y1==0,NA)
ds$cT4<-replace(ds$cT4, ds$y1==0,NA)
ds$uT4<-replace(ds$uT4, ds$y2==0,NA)
ds$cT4<-replace(ds$cT4, ds$y2==0,NA)
ds$uT4<-replace(ds$uT4, ds$y3==0,NA)
ds$cT4<-replace(ds$cT4, ds$y3==0,NA)
ds$z4 <- ifelse(ds$trt == 1, + 0.300*ds$age + 1*ds$gender + 0.65*ds$uT0 + 0.00027*ds$cT0 + 1*ds$trt, 1)
ds$z4 <- ifelse(ds$trt == 0, - 0.300*ds$age - 1*ds$gender - 0.65*ds$uT0 - 0.00010*ds$cT0 + 1*ds$trt, ds$z4)
ds$pr4a = exp(-ds$z4)
ds$pr4b = 1 + ds$pr4a
ds$pr4c = 1/ ds$pr4b
ds$y4 = rbinom(600,1,ds$pr4c)
ds$uT4<-replace(ds$uT4, ds$trt==1 & ds$y4==0 & ds$uT4 >= ds$mediantotalcost,NA)
ds$cT4<-replace(ds$cT4, ds$trt==0 & ds$y4==0 & ds$cT4 <= ds$medianQALYs,NA)
descr(ds$y4)
ds$M4 <- as.integer(complete.cases(ds$cT4))
descr(ds$M4)
```

```
 ds$M <- as.integer(complete.cases(ds))
 descr(ds$M)
 ds <- as.data.frame(ds)
 ds <- subset(ds,select =c(“id”,”trt”,”age”,”gender”,”uT0”,”cT0”,”uT1”,”cT1”,”uT2”,”cT2”,”uT3”,”cT3”,”uT4”,”cT4”))
 }
test <- lapply(data,MISSING50)
for (i in 1:2000) {
 write.dta(test[[i]], file = paste0(“C:/MISSING50/dataset”,i,”.dta”)}
```

SUPPLEMENTARY MATERIAL 3

/* Stata code - LONGITUDINAL LINEAR MIXED-MODEL (LLM) */

```
 clear
 set more off
 cd “C:\MISSINGX”
 local n = 2000

forvalues j = 1(1)`n’{
 local y = `j’
 use “dataset`y’”, clear

 //keep baseline values for costs and utilities in wide format
 gen utility_b = uT0
 gen cost_b = cT0

 //reshape from wide to long creating a new variable - time - that indicates time period
 quietly reshape long cT uT, i(id) j(time 0 1 2 3 4)
 rename cT costs
 rename uT utilities

 //LLM FOR UTILITIES
 mixed utilities i.trt##i.time i.time##c.utility_b i.time##c.age i.time##i.gender ||id :
 mat betaCE = e(b) /* extract matric of betas */
 mat vari = e(V) /* extract matrix of variances */
 mat obs = e(N_g)
 gen obs = obs[1,1] /* extract number of observations used in the model */
 gen Za = 1.95996

 //gen variables for utility difference at each time point adjusted for baseline
 gen utility_diff1 = betaCE[1,14]
 gen utility_diff2 = betaCE[1,15]
 gen utility_diff3 = betaCE[1,16]
 gen utility_diff4 = betaCE[1,17]

 //calculate QALY difference
  gen QALY_diff = (0.5*(utility_diff1+utility_diff1)*(3/12)) + (0.5*(utility_diff1+utility_diff2)*(3/12)) + (0.5*(utility_diff2+utility_diff3)*(3/12)) + (0.5*(utility_diff3+utility_diff4)*(3/12))

 //gen variables for variance utility difference at each time point
 gen varu1 = vari[14,14]
 gen varu2 = vari[15,15]
 gen varu3 = vari[16,16]
 gen varu4 = vari[17,17]

 //calculate variance of QALY difference
 gen QALY_var = ((0.25*(varu1+varu1)) + (0.25*(varu1+varu2)) + (0.25*(varu2+varu3)) + (0.25*(varu3+varu4)))/4
```

```
//calculate SE of QALY difference
gen SE_QALY_diff = sqrt(QALY_var)

//estimate CI around QALY difference
gen LL_QALY = QALY_diff - Za*SE_QALY_diff
gen UL_QALY = QALY_diff + Za*SE_QALY_diff

//LLM FOR COSTS
mixed costs i.trt##i.time i.time##c.cost_b i.time##c.age i.time##i.gender ||id :
mat betaCEc = e(b) /* extract matrix of regression coefficients */
mat varic = e(V) /* extract matrix of variances */
mat obsc = e(N_g)
gen obsc = obsc[1,1] /* extract number of observations used in the model */

//gen variables for cost difference at each time point adjusted for baseline costs
gen cost_diff1 = betaCEc[1,14]
gen cost_diff2 = betaCEc[1,15]
gen cost_diff3 = betaCEc[1,16]
gen cost_diff4 = betaCEc[1,17]

//calculate marginal cost difference
gen cost_diff = cost_diff1 + cost_diff2 + cost_diff3 + cost_diff4

//gen variables for variance of cost difference at each time point
gen varc1 = varic[14,14]
gen varc2 = varic[15,15]
gen varc3 = varic[16,16]
gen varc4 = varic[17,17]

//calculate SE for marginal cost difference
gen SEc1 = sqrt(varc1)
gen SEc2 = sqrt(varc2)
gen SEc3 = sqrt(varc3)
gen SEc4 = sqrt(varc4)
gen SE_cost_diff = SEc1 + SEc2 + SEc3 + SEc4

//estimate CI around marginal cost difference
gen LL_costs = (cost_diff - Za*SE_cost_diff)
gen UL_costs = (cost_diff + Za*SE_cost_diff)

//calculate covariance costs and QALY
cor costs utilities if trt == 0
mat cor = r(rho)
gen cor_control = cor[1,1]

cor costs utilities if trt == 1
mat cor = r(rho)
gen cor_intervention = cor[1,1]
```

```
 tabstat costs, stats(sd) save by(trt)
 mat sd = r(Stat1)
 gen Costs_SD_control = sd[1,1]

 tabstat costs, stats(sd) save by(trt)
 mat sd = r(Stat2)
 gen Costs_SD_intervention = sd[1,1]

 tabstat utilities, stats(sd) save by(trt)
 mat sd = r(Stat1)
 gen QALY_SD_control = sd[1,1]

 tabstat utilities, stats(sd) save by(trt)
 mat sd = r(Stat2)
 gen QALY_SD_intervention = sd[1,1]

 gen QALY_control_se = QALY_SD_control/sqrt(obs)
 gen QALY_int_se = QALY_SD_intervention/sqrt(obs)

 gen Costs_control_se = Costs_SD_control/sqrt(obs)
 gen Costs_int_se = Costs_SD_intervention/sqrt(obs)
 gen se_QALY = sqrt(QALY_int_se^2+QALY_control_se^2)
 gen se_Costs = sqrt(Costs_int_se^2+Costs_control_se^2)
 gen cor_cost_effect_diffs = ((cor_intervention*QALY_int_se*Costs_int_se)+(cor_control*QALY_control_se*Costs_control_se))/(se_QALY*se_Costs)
 gen cov = cor_cost_effect_diffs*SE_QALY_diff*SE_cost_diff
 save “LLM\postboots`y’”, replace
}
```

/* MULTIPLE IMPUTATION + LONGITUDINAL LINEAR MIXED-MODEL (MI-LLM) */

```
 clear
 set more off
 cd “C:\MISSINGX”
 local n = 2000

forvalues j = 1(1)`n’{
 local y = `j’
 use “dataset`y’”, clear

// MULTIPLE IMPUTATION MODEL
 mi set flong
 mi register regular age gender trt uT0 cT0
 mi register imputed cT1 cT2 cT3 cT4 uT1 uT2 uT3 uT4
 quietly mi impute chained (pmm, knn(5)) cT1 cT2 cT3 cT4 uT1 uT2 uT3 uT4 = age gender uT0 cT0, by(trt) replace add(10) rseed(`k’)
 save «C:\MISSINGX\MI-LLM\dataset`y’_imp», replace
}

clear
set more off
cd “C:\MISSING\MI-LLM”
```

```
local n = 2000
forvalues j = 1(1)`n’{
 local y = `j’
 use “dataset`y’_imp”, clear

 //keep baseline values for costs and utilities in wide format
 gen utility_b = uT0
 gen cost_b = cT0

 //reshape from wide to long creating a new variable - time - that indicates time period
 rename (cT0 cT1 cT2 cT3 cT4) (cost0 cost1 cost2 cost3 cost4)
 rename (uT0 uT1 uT2 uT3 uT4) (utility0 utility1 utility2 utility3 utility4)
 quietly mi reshape long cost utility, i(id) j(time 0 1 2 3 4)

 //LLM FOR UTILITIES performed in each on of the imputed datasets
 quietly mi estimate: mixed utility i.trt##i.time i.time##c.utility_b i.time##c.age i.time##i.gender ||id :
 mat betaCE = e(b_mi) /* extract matric of betas */
 mat vari = e(V_mi) /* extract matrix of variances */
 mat obs = e(N_g_mi) /* extract number of observations used in the model */
 gen obs = obs[1,1]
 gen loss_eff = e(fmi_max_mi)/e(M_mi)
 gen Za = 1.95996

 //gen variables for utility difference at each time point adjusted for baseline */
 gen utility_diff1 = betaCE[1,14]
 gen utility_diff2 = betaCE[1,15]
 gen utility_diff3 = betaCE[1,16]
 gen utility_diff4 = betaCE[1,17]

 //calculate QALY difference
  gen QALY_diff = (0.5*(utility_diff1+utility_diff1)*(3/12)) + (0.5*(utility_diff1+utility_diff2)*(3/12)) + (0.5*(utility_diff2+utility_diff3)*(3/12)) + (0.5*(utility_diff3+utility_diff4)*(3/12))

 //gen variables for variance utility difference at each time point
 gen varu1 = vari[14,14]
 gen varu2 = vari[15,15]
 gen varu3 = vari[16,16]
 gen varu4 = vari[17,17]

//calculate variance of QALY difference
 gen QALY_var = ((0.25*(varu1+varu1)) + (0.25*(varu1+varu2)) + (0.25*(varu2+varu3)) + (0.25*(varu3+varu4)))/4

 //calculate SE of QALY difference
 gen SE_QALY_diff = sqrt(QALY_var)

 //estimate CI around QALY difference
 gen LL_QALY = QALY_diff - Za*SE_QALY_diff
 gen UL_QALY = QALY_diff + Za*SE_QALY_diff
```

```
//LLM FOR COSTS performed in each on of the imputed datasets
quietly mi estimate: mixed cost i.trt##i.time i.time##c.cost_b i.time##c.age i.time##i.gender ||id :
mat betaCEc = e(b_mi) /* extract matric of betas */
mat varic = e(V_mi) /* extract matrix of variances */
mat obsc = e(N_g_mi) /* extract number of observations used in the model */
gen obsc = obsc[1,1]
gen loss_effc = e(fmi_max_mi)/e(M_mi)

//gen variables for cost difference at each time point adjusted for baseline costs
gen cost_diff1 = betaCEc[1,14]
gen cost_diff2 = betaCEc[1,15]
gen cost_diff3 = betaCEc[1,16]
gen cost_diff4 = betaCEc[1,17]

//calculate marginal cost difference
gen cost_diff = cost_diff1 + cost_diff2 + cost_diff3 + cost_diff4

//gen variables for variance of cost difference at each time point
gen varc1 = varic[14,14]
gen varc2 = varic[15,15]
gen varc3 = varic[16,16]
gen varc4 = varic[17,17]

//calculate SE for marginal cost difference
gen SEc1 = sqrt(varc1)
gen SEc2 = sqrt(varc2)
gen SEc3 = sqrt(varc3)
gen SEc4 = sqrt(varc4)
gen SE_cost_diff = SEc1 + SEc2 + SEc3 + SEc4

//estimate CI around marginal cost difference
gen LL_costs = (cost_diff - Za*SE_cost_diff)
gen UL_costs = (cost_diff + Za*SE_cost_diff)

//calculate covariance costs and QALY
cor cost utility if trt == 0
mat cor = r(rho)
gen cor_control = cor[1,1]

cor cost utility if trt == 1
mat cor = r(rho)
gen cor_intervention = cor[1,1]

tabstat cost, stats(sd) save by(trt)
mat sd = r(Stat1)
gen Costs_SD_control = sd[1,1]

tabstat cost, stats(sd) save by(trt)
mat sd = r(Stat2)
gen Costs_SD_intervention = sd[1,1]
```

```
 tabstat utility, stats(sd) save by(trt)
 mat sd = r(Stat1)
 gen QALY_SD_control = sd[1,1]

 tabstat utility, stats(sd) save by(trt)
 mat sd = r(Stat2)
 gen QALY_SD_intervention = sd[1,1]

 gen QALY_control_se = QALY_SD_control/sqrt(obs)
 gen QALY_int_se = QALY_SD_intervention/sqrt(obs)

 gen Costs_control_se = Costs_SD_control/sqrt(obs)
 gen Costs_int_se = Costs_SD_intervention/sqrt(obs)

 gen se_QALY = sqrt(QALY_int_se^2+QALY_control_se^2)
 gen se_Costs = sqrt(Costs_int_se^2+Costs_control_se^2)

 gen cor_cost_effect_diffs = ((cor_intervention*QALY_int_se*Costs_int_se)+(cor_control*QALY_control_se*Costs_control_se))/(se_QALY*se_Costs)

 gen cov = cor_cost_effect_diffs*SE_QALY_diff*SE_cost_diff

 save “C:\MISSINGX\MI-LLM\postboots`y’”, replace
}
```

/* MEAN IMPUTATION + LONGITUDINAL LINEAR MIXED-MODEL (M-LLM) */

```
clear
set more off
cd «C:\MISSINGX»
local n = 2000

forvalues j = 1(1)`n’{
 local y = `j’
 use «dataset`y’», clear

// MEAN IMPUTATION BY TREATMENT GROUP
 bysort trt: egen mean_uT1 = mean(uT1)
 bysort trt: egen mean_uT2 = mean(uT2)
 bysort trt: egen mean_uT3 = mean(uT3)
 bysort trt: egen mean_uT4 = mean(uT4)

 replace uT1 = mean_uT1 if missing(uT1)
 replace uT2 = mean_uT2 if missing(uT2)
 replace uT3 = mean_uT3 if missing(uT3)
 replace uT4 = mean_uT4 if missing(uT4)

 bysort trt: egen mean_cT1 = mean(cT1)
 bysort trt: egen mean_cT2 = mean(cT2)
 bysort trt: egen mean_cT3 = mean(cT3)
 bysort trt: egen mean_cT4 = mean(cT4)

 replace cT1 = mean_cT1 if missing(cT1)
 replace cT2 = mean_cT2 if missing(cT2)
 replace cT3 = mean_cT3 if missing(cT3)
 replace cT4 = mean_cT4 if missing(cT4)
//keep baseline values for costs and utilities in wide format
 gen utility_b = uT0
 gen cost_b = cT0

 //reshape from wide to long creating a new variable - time - that indicates time period
 quietly reshape long cT uT, i(id) j(time 0 1 2 3 4)
 rename cT costs
 rename uT utilities

 //LLM FOR UTILITIES
 mixed utilities i.trt##i.time i.time##c.utility_b i.time##c.age i.time##i.gender ||id :
 mat betaCE = e(b) /* extract matric of betas */
 mat vari = e(V) /* extract matrix of variances */
 mat obs = e(N_g)
 gen obs = obs[1,1] /* extract number of observations used in the model */
 gen Za = 1.95996

 //gen variables for utility difference at each time point adjusted for baseline
 gen utility_diff1 = betaCE[1,14]
 gen utility_diff2 = betaCE[1,15]
 gen utility_diff3 = betaCE[1,16]
 gen utility_diff4 = betaCE[1,17]

 //calculate QALY difference
  gen QALY_diff = (0.5*(utility_diff1+utility_diff1)*(3/12)) + (0.5*(utility_diff1+utility_diff2)*(3/12)) + (0.5*(utility_
diff2+utility_diff3)*(3/12)) + (0.5*(utility_diff3+utility_diff4)*(3/12))

 //gen variables for variance utility difference at each time point
 gen varu1 = vari[14,14]
 gen varu2 = vari[15,15]
 gen varu3 = vari[16,16]
 gen varu4 = vari[17,17]

 //calculate variance of QALY difference
 gen QALY_var = ((0.25*(varu1+varu1)) + (0.25*(varu1+varu2)) + (0.25*(varu2+varu3)) + (0.25*(varu3+varu4)))/4

 //calculate SE of QALY difference
 gen SE_QALY_diff = sqrt(QALY_var)

 //estimate CI around QALY difference
 gen LL_QALY = QALY_diff - Za*SE_QALY_diff
 gen UL_QALY = QALY_diff + Za*SE_QALY_diff

 //LLM FOR COSTS
 mixed costs i.trt##i.time i.time##c.cost_b i.time##c.age i.time##i.gender ||id :
 mat betaCEc = e(b) /* extract matrix of regression coefficients */
 mat varic = e(V) /* extract matrix of variances */
mat obsc = e(N_g)
gen obsc = obsc[1,1] /* extract number of observations used in the model */

//gen variables for cost difference at each time point adjusted for baseline costs
gen cost_diff1 = betaCEc[1,14]
gen cost_diff2 = betaCEc[1,15]
gen cost_diff3 = betaCEc[1,16]
gen cost_diff4 = betaCEc[1,17]

//calculate marginal cost difference
gen cost_diff = cost_diff1 + cost_diff2 + cost_diff3 + cost_diff4

//gen variables for variance of cost difference at each time point
gen varc1 = varic[14,14]
gen varc2 = varic[15,15]
gen varc3 = varic[16,16]
gen varc4 = varic[17,17]

//calculate SE for marginal cost difference
gen SEc1 = sqrt(varc1)
gen SEc2 = sqrt(varc2)
gen SEc3 = sqrt(varc3)
gen SEc4 = sqrt(varc4)
gen SE_cost_diff = SEc1 + SEc2 + SEc3 + SEc4

//estimate CI around marginal cost difference
gen LL_costs = (cost_diff - Za*SE_cost_diff)
gen UL_costs = (cost_diff + Za*SE_cost_diff)

//calculate covariance costs and QALY
cor costs utilities if trt == 0
mat cor = r(rho)
gen cor_control = cor[1,1]

cor costs utilities if trt == 1
mat cor = r(rho)
gen cor_intervention = cor[1,1]

tabstat costs, stats(sd) save by(trt)
mat sd = r(Stat1)
gen Costs_SD_control = sd[1,1]

tabstat costs, stats(sd) save by(trt)
mat sd = r(Stat2)
gen Costs_SD_intervention = sd[1,1]

tabstat utilities, stats(sd) save by(trt)
mat sd = r(Stat1)
gen QALY_SD_control = sd[1,1]
```

```
 tabstat utilities, stats(sd) save by(trt)
 mat sd = r(Stat2)
 gen QALY_SD_intervention = sd[1,1]

 gen QALY_control_se = QALY_SD_control/sqrt(obs)
 gen QALY_int_se = QALY_SD_intervention/sqrt(obs)

 gen Costs_control_se = Costs_SD_control/sqrt(obs)
 gen Costs_int_se = Costs_SD_intervention/sqrt(obs)

 gen se_QALY = sqrt(QALY_int_se^2+QALY_control_se^2)
 gen se_Costs = sqrt(Costs_int_se^2+Costs_control_se^2)

  gen  cor_cost_effect_diffs  =  ((cor_intervention*QALY_int_se*Costs_int_se)+(cor_control*QALY_control_se*Costs_control_se))/(se_QALY*se_Costs)

 gen cov = cor_cost_effect_diffs*SE_QALY_diff*SE_cost_diff

 save “M-LLM\postboots`y’”, replace
}
```

/*SEEMINGLY UNRELATED REGRESSION COMPLETE CASE ANALYSIS (SUR)*/

```
 clear
 set more off
 cd “C:\MISSINGX”
 local n = 2000

forvalues j = 1(1)`n’{
 local y = `j’
 use “dataset`y’”, clear

 gen QALY = (0.5*(uT1+uT1)*(3/12)) + (0.5*(uT1+uT2)*(3/12)) + (0.5*(uT2+uT3)*(3/12)) + (0.5*(uT3+uT4)*(3/12))
 gen Tcosts = cT1 + cT2 + cT3 + cT4

 quietly sureg (Tcosts=trt cT0 age gender) (QALY=trt uT0 age gender)
 mat betaCE= e(b) /* extracT the matrix of regression coefficients */
 mat vari = e(V) /* extracT matrix of variances */
 gen obs = e(N) /* extracT number of observations used in the model */

 gen cost_diff = betaCE[1,1] /* generate cost difference */
 gen QALY_diff = betaCE[1,6] /* generate QALY difference */

 gen cost_var = vari[1,1] /* generate variance of cost difference */
 gen QALY_var = vari[6,6] /* generate variance of QALY difference */
 gen cov = vari[1,6] /* generate covariance cost and QALY difference */

 gen SE_cost_diff = sqrt(cost_var)
 gen SE_QALY_diff = sqrt(QALY_var)
 gen Za = 1.95996
 gen LL_costs = cost_diff - Za*SE_cost_diff /* lower CI cost difference */
 gen UL_costs = cost_diff + Za*SE_cost_diff /* upper CI cost difference */

 gen LL_QALY = QALY_diff - Za*SE_QALY_diff /* lower CI QALY difference */
 gen UL_QALY = QALY_diff + Za*SE_QALY_diff /* upper CI QALY difference */

 save “SUREG\postboots`y’”, replace
}
```

/* MULTIPLE IMPUTATION + SEEMINGLY UNRELATE REGRESSION (MI-SUR) */

```
 clear
 set more off
 cd “C:\MISSINGX”
 local n = 2000

forvalues k=1(1)`n’{
 local y = `k’
 use “dataset`y’”, clear

// MULTIPLE IMPUTATION MODEL
 mi set flong
 mi register regular age gender trt uT0 cT0
 mi register imputed cT1 cT2 cT3 cT4 uT1 uT2 uT3 uT4
 quietly mi impute chained (pmm, knn(5)) cT1 cT2 cT3 cT4 uT1 uT2 uT3 uT4 = age gender uT0 cT0, by(trt) replace
add(10) rseed(`k’)

 save “C:\MISSINGX\MI-LLM\dataset`y’_imp”, replace
 save “C:\MISSINGX\MI-SUREG\dataset`y’_imp”, replace

 gen QALY = (0.5*(uT1+uT1)*(3/12)) + (0.5*(uT1+uT2)*(3/12)) + (0.5*(uT2+uT3)*(3/12)) + (0.5*(uT3+uT4)*(3/12))
 gen Tcosts = cT1 + cT2 + cT3 + cT4

 quietly mi estimate, cmdok: sureg (Tcosts=trt cT0 age gender) (QALY=trt uT0 age gender)
 gen obs = e(N_mi)
 gen loss_eff = e(fmi_max_mi)/e(M_mi)

 mat betaCE= e(b_mi) /* extract the matrix of regression coefficients */
 mat vari = e(V_mi) /* extract matrix of variances */

 gen cost_diff = betaCE[1,1] /* generate cost difference */
 gen QALY_diff = betaCE[1,6] /* generate QALY difference */

 gen cost_var = vari[1,1] /* generate variance of cost difference */
 gen QALY_var = vari[6,6] /* generate variance of QALY difference */
 gen cov = vari[1,6] /* generate covariance cost and QALY difference */
 gen SE_cost_diff = sqrt(cost_var)
 gen SE_QALY_diff = sqrt(QALY_var)

 gen Za = 1.95996
 gen LL_costs = cost_diff - Za*SE_cost_diff /* lower CI cost difference */
 gen UL_costs = cost_diff + Za*SE_cost_diff /* upper CI cost difference */

 gen LL_QALY = QALY_diff - Za*SE_QALY_diff /* lower CI QALY difference */
 gen UL_QALY = QALY_diff + Za*SE_QALY_diff /* upper CI QALY difference */

 save “C:\MISSINGX\MI-SUREG\postboots`y’”, replace
 }
```

/* MEAN IMPUTATION + SEEMENGLY UNRELATED REGRESSION (M-SUR) */

```
 clear
 set more off
 cd “C:\MISSINGX”
 local n = 2000

forvalues j = 1(1)`n’{
 local y = `j’
 use “dataset`y’”, clear

// MEAN IMPUTATION BY TREATMENT GROUP
 bysort trt: egen mean_uT1 = mean(uT1)
 bysort trt: egen mean_uT2 = mean(uT2)
 bysort trt: egen mean_uT3 = mean(uT3)
 bysort trt: egen mean_uT4 = mean(uT4)

 replace uT1 = mean_uT1 if missing(uT1)
 replace uT2 = mean_uT2 if missing(uT2)
 replace uT3 = mean_uT3 if missing(uT3)
 replace uT4 = mean_uT4 if missing(uT4)

 bysort trt: egen mean_cT1 = mean(cT1)
 bysort trt: egen mean_cT2 = mean(cT2)
 bysort trt: egen mean_cT3 = mean(cT3)
 bysort trt: egen mean_cT4 = mean(cT4)

 replace cT1 = mean_cT1 if missing(cT1)
 replace cT2 = mean_cT2 if missing(cT2)
 replace cT3 = mean_cT3 if missing(cT3)
 replace cT4 = mean_cT4 if missing(cT4)

 gen QALY = (0.5*(uT1+uT1)*(3/12)) + (0.5*(uT1+uT2)*(3/12)) + (0.5*(uT2+uT3)*(3/12)) + (0.5*(uT3+uT4)*(3/12))
 gen Tcosts = cT1 + cT2 + cT3 + cT4

 quietly sureg (Tcosts=trt cT0 age gender) (QALY=trt uT0 age gender)
 mat betaCE= e(b) /* extract the matrix of regression coefficients */
 mat vari = e(V) /* extract matrix of variances */
 gen obs = e(N) /* extract number of observations used in the model */

 gen cost_diff = betaCE[1,1] /* generate cost difference */
 gen QALY_diff = betaCE[1,6] /* generate QALY difference */

 gen cost_var = vari[1,1] /* generate variance of cost difference */
 gen QALY_var = vari[6,6] /* generate variance of QALY difference */
 gen cov = vari[1,6] /* generate covariance cost and QALY difference */

 gen SE_cost_diff = sqrt(cost_var)
 gen SE_QALY_diff = sqrt(QALY_var)

 gen Za = 1.95996
 gen LL_costs = cost_diff - Za*SE_cost_diff /* lower CI cost difference */
 gen UL_costs = cost_diff + Za*SE_cost_diff /* upper CI cost difference */

 gen LL_QALY = QALY_diff - Za*SE_QALY_diff /* lower CI QALY difference */
 gen UL_QALY = QALY_diff + Za*SE_QALY_diff /* upper CI QALY difference */

 save "M-SUREG\postboots`y'", replace
}
```

Supplementary Material 4

Empirical dataset 1

Data from two pragmatic randomized controlled trials were used in addition to the simulated data. In the first trial (empirical dataset 1), the cost-effectiveness of early rehabilitation after lumbar disc surgery was compared to no referral.[18] For the current study, utility values collected at baseline, 12, and 26 weeks and costs collected at 6, 12, and 26 weeks were used. For all scenarios, mean imputation was used to impute missing values at baseline.[5] Of the 169 participants used in our study, 13% (n=22) had missing cost and/or utility data at one or more follow-up time points. Stepwise backwards regression models with p<0.05, were used to identify baseline variables that were predictive of the missingness of data and/or the cost-effectiveness outcomes. The identified variables were added to the imputation model as auxiliary variables (i.e. age, level of education, utility values, Oswestry Disability Index [ODI], pain intensity, Örebro Musculoskeletal Pain Screening Questionnaire [OMPSQ], and the credibility and expectancy surgery [CEQ])[18]. Missing cost and utility data were imputed using Multivariate Imputation by Chained Equations (MICE; FCS-standard)[33] with PMM, stratified by treatment group.[34] Ten datasets were imputed to guarantee a loss of efficiency <0.05. SUR and LLM were then fitted to the imputed data. The LLM analysis model included all auxiliary variables as, in doing so, it should lead to similar results when compared to MI-LL.[5] The other LMM models (i.e., M-LLM and MI-LLM) and SUR models (i.e., SUR-CCA, M-SUR, and MI-SUR) did not include auxiliary variables, and only were corrected for confounders (i.e. baseline utility values, ODI, OMPSQ, and CEQ).

Longitudinal Linear Mixed-model analysis (LLM) – Two separate LLMs were performed, including one for costs and one for utility values:

$$Costs_{ij} = \beta_{1c}time_j + \beta_{2c}trt_i + \beta_{3c}time_j trt_i + \beta_{4c}uT0_i + \beta_{5c}NRS_i + \beta_{6c}ODI_i + \beta_{7c}CEQ_i + \omega_{ci} + \varepsilon_{cij},$$

$$Utility_{ij} = \beta_{1u}time_j + \beta_{2u}trt_i + \beta_{3u}time_j trt_i + \beta_{4u}uT0_i + \beta_{5u}age_i + \beta_{6u}educ_i + \beta_{7u}OMPSQ_i + \beta_{8u}CEQ_i + \omega_{ui} + \varepsilon_{uij},$$

$$\omega_i \sim Normal(0, \sigma_\omega^2),\ \varepsilon_{ij} \sim Normal(0, \sigma_\varepsilon^2)$$

where $Costs_{ij}$ and $Utility_{ij}$ represent the cost and utility values of subject i (i = 1, …, N=169) at time point j (j = 1, …, 3). The model parameters include the intercept β_1 and the coefficients $\beta_{2,\dots,}\beta_n$ of covariates including $time_j$ as an interaction term for the treatment effect. ω_{ci} and ω_{ui} represent the random intercepts and ε_{ij} and ε_{uij} represent the error term for a patient i at each time point j for $Costs$ and $Utility$, respectively. Both ω_i and ε_{ij} follow a normal distribution.

Mean Imputation combined with LLM (M-LLM) – In this strategy, missing cost and utility values were replaced by the mean values from the available cases at each time point (i.e., unconditional mean imputation).[5] Subsequently, two separate LLMs were fitted but not including the variables associated with missingness:

$$Costs_{ij} = \beta_{1c}time_j + \beta_{2c}trt_i + \beta_{3c}time_j trt_i + \beta_{4c}uT0_i + \beta_{5c}ODI_i + \beta_{6c}CEQ_i + \omega_{ci} + \varepsilon_{cij},$$

$$Utility_{ij} = \beta_{1u}time_j + \beta_{2u}trt_i + \beta_{3u}time_j trt_i + \beta_{4u}uT0_i + \beta_{5u}OMPSQ_i + \beta_{6u}CEQ_i + \omega_{ui} + \varepsilon_{uij},$$

$$\omega_i \sim Normal(0, \sigma_\omega^2),\ \varepsilon_{ii} \sim Normal(0, \sigma_\varepsilon^2)$$

Multiple Imputation combined with LLM (MI-LLM) – In this strategy, missing cost and utility values were first imputed using Multivariate Imputation by Chained Equations (MICE; FCS-standard)[33] with Predictive Mean Matching (PMM) by treatment group as outlined above.[34] Subsequently, two separate LLMs were fitted but not including the variables associated with missingness as show in the M-LLM strategy.

Seemingly Unrelated Regressions − Complete Case Analysis (SUR-CCA) – In this strategy, all subjects with missing values were deleted from the datasets (i.e. a complete case analysis). Then, total costs and QALYs were calculated by adding costs at each time point and using the area under the curve method, respectively.[23] Total cost and QALY differences between treatment groups were estimated using seemingly unrelated regressions (SUR). With SUR two regression equations are modelled simultaneously (i.e., one for total costs and one for QALY), while correcting for their possible correlation through correlated error terms:[16,35]

$$Costs_i = \beta_{0c} + \beta_{1c}trt_i + \beta_{2c}ODI_i + \beta_{3c}CEQ_i + \varepsilon_{ci}$$

$$QALYs_i = \beta_{0q} + \beta_{1q}trt_i + \beta_{2q}uT0_i + \beta_{3q}OMPSQ_i + \beta_{4q}CEQ_i + \varepsilon_{qi}$$

$$\begin{pmatrix}\varepsilon_{ci}\\ \varepsilon_{qi}\end{pmatrix} \sim Normal\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_c^2 & \sigma_{cq}\\ \sigma_{cq} & \sigma_q^2\end{pmatrix}\right)$$

Mean Imputation combined with SUR (M-SUR) – In this strategy, missing cost and utility data were replaced by the mean values from the available cases at each time point (i.e., unconditional mean imputation).[5] Subsequently, total costs and QALYs were calculated, and SUR analyses were performed as outlined under SUR.

Multiple Imputation combined with SUR (MI-SUR) – In this strategy, missing cost and utility data were first imputed using Multivariate Imputation by Chained Equations (MICE) as outlined above. Then, total costs and QALYs were calculated and a SUR was fitted per imputed dataset as outlined under SUR, after which pooled estimates were obtained using Rubin's rules[33]. The SUR model as fitted as shown by SUR equations above.

Empirical dataset 2

In the second trial (empirical dataset 2), the cost-effectiveness of an interpersonal psychotherapy for older adults with major depression was compared to care as usual (i.e., control). For this study, utility values collected at baseline, 6, and 12 months and costs collected at 2, 6, and 12 months were used,[19] Mean imputation was used to impute missing values at baseline.[5] Of the 143 participants, 68% (n=98) of cost and utility data were missing at one or more follow-up time points. Stepwise backwards regression models with p<0.05, were used to identify baseline variables that were predictive of the missingness of data and/or the cost-effectiveness outcomes. The identified variables were added to the imputation model as auxiliary variables (i.e., age, activity daily living [ADL], utility values, alcohol-induced disorder, and mental health problems utility values, marital status, and household composition).[19] Missing cost and utility data were imputed using Multivariate Imputation by Chained Equations (MICE; FCS-standard)[33] with PMM by treatment group.[34] Twenty datasets were imputed to guarantee a loss of efficiency <0.05. SUR and LLM were then fitted to the imputed data. The LLM analysis model included all auxiliary variables.[5] The other LMM models (i.e., M-LLM and MI-LLM) and SUR models (i.e., SUR-CCA, M-SUR, and MI-SUR) did not include auxiliary variables and only were corrected for confounders (i.e., baseline utility values, marital status, and household composition).

Longitudinal Linear Mixed-model analysis (LLM) – Two separate LLMs were performed, including one for costs and one for utility values:

$$Costs_{ij} = \beta_{1c}time_j + \beta_{2c}trt_i + \beta_{3c}time_j trt_i + \beta_{4c}uT0_i + \beta_{5c}ALC_i + \beta_{6c}MS_i + \beta_{7c}SF36mh_i + \omega_{ci} + \varepsilon_{cij},$$

$$Utility_{ij} = \beta_{1u}time_j + \beta_{2u}trt_i + \beta_{3u}time_j trt_i + \beta_{4u}uT0_i + \beta_{5u}age_i + \beta_{6u}ADL_i + \beta_{7u}HC_i + \beta_{8u}MS_i + \omega_{ui} + \varepsilon_{uij},$$

$$\omega_i \sim Normal(0, \sigma_\omega^2),\ \varepsilon_{ij} \sim Normal(0, \sigma_\varepsilon^2)$$

where $Costs_{ij}$ and $Utility_{ij}$ represent the cost and utility values of subject *i* (*i* = 1, …, N=143) at time point j (*j* = 1, …, 3). The model parameters include the intercept β_1 and the coefficients $\beta_{2,...,}\beta_n$ of covariates including $time_j$ as an interaction term for the treatment effect. ω_{ci} and ω_{ui} represent the random intercepts and ε_{cij} and ε_{uij} represent the error terms for a patient *i* at each time point *j* for *Costs* and *Utility*, respectively. Both ε_{ij} and follow a normal distribution.

Mean Imputation combined with LLM (M-LLM) – In this strategy, missing cost and utility values were replaced by the mean values from the available cases at each time point (i.e., unconditional mean imputation).[5] Subsequently, two separate LLMs were fitted but not including the variables associated with missingness:

$$Costs_{ij} = \beta_{1c}time_j + \beta_{2c}trt_i + \beta_{3c}time_j trt_i + \beta_{4c}uT0_i + \beta_{5c}MS_i + \omega_{ci} + \varepsilon_{cij},$$

$$Utility_{ij} = \beta_{1u}time_j + \beta_{2u}trt_i + \beta_{3u}time_j trt_i + \beta_{4u}uT0_i + \beta_{5u}MS_i + \beta_{6u}HC_i + \omega_{ui} + \varepsilon_{uij},$$

$$\omega_i \sim Normal(0, \sigma_\omega^2),\ \varepsilon_{ij} \sim Normal(0, \sigma_\varepsilon^2)$$

Multiple Imputation combined with LLM (MI-LLM) – In this strategy, missing cost and utility values were first imputed using Multivariate Imputation by Chained Equations (MICE; FCS-standard)[33] with Predictive Mean Matching (PMM) by treatment group as outlined above.[34] Subsequently, two separate LLMs were fitted but not including the variables associated with missingness as show in the M-LLM strategy.

Seemingly Unrelated Regressions − Complete Case Analysis (SUR-CCA) – In this strategy, all subjects with missing values were deleted from the datasets (i.e. a complete case analysis). Then, total costs and QALYs were calculated by adding costs at each time point and using the area under the curve method, respectively.[23] Total cost and QALY differences between treatment groups were estimated using seemingly unrelated regressions (SUR). With SUR two regression equations are modelled simultaneously (i.e., one for total costs and one for QALY), while correcting for their possible correlation through correlated error terms:[16,35]

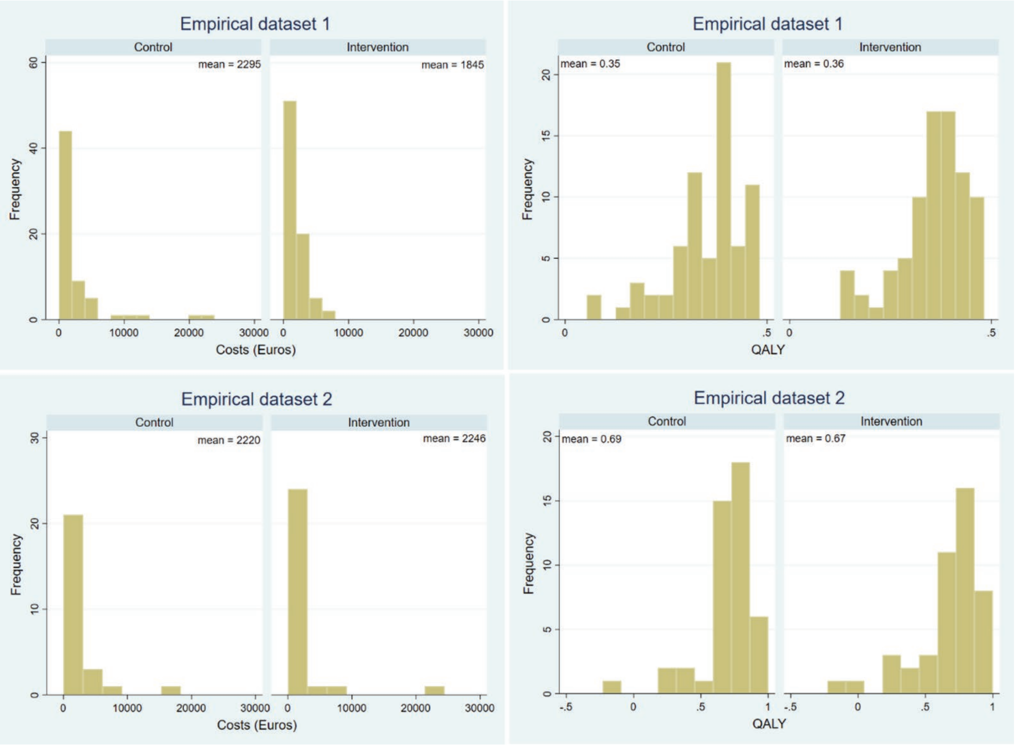
$$Costs_i = \beta_{0c} + \beta_{1c}trt_i + \beta_{42}uT0_i + \beta_{3c}MS_i + \varepsilon_{ci}$$

$$QALYs_i = \beta_{0q} + \beta_{1q}trt_i + \beta_{2q}uT0_i + \beta_{3q}MS_i + \beta_{4q}HC_i + \varepsilon_{qi}$$

$$\begin{pmatrix}\varepsilon_{ci}\\ \varepsilon_{qi}\end{pmatrix} \sim Normal\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_c^2 & \sigma_{cq}\\ \sigma_{cq} & \sigma_q^2\end{pmatrix}\right)$$

Mean Imputation combined with SUR (M-SUR) – In this strategy, missing cost and utility data were replaced by the mean values from the available cases at each time point (i.e., unconditional mean imputation)[5]. Subsequently, total costs and QALYs were calculated, and SUR analyses were performed as outlined under SUR.

Multiple Imputation combined with SUR (MI-SUR) – In this strategy, missing cost and utility data were first imputed using Multivariate Imputation by Chained Equations (MICE) as outlined above. Then, total costs and QALYs were calculated and a SUR was fitted per imputed dataset as outlined under SUR, after which pooled estimates were obtained using Rubin's rules.[33] The SUR model as fitted as shown by SUR equations above.

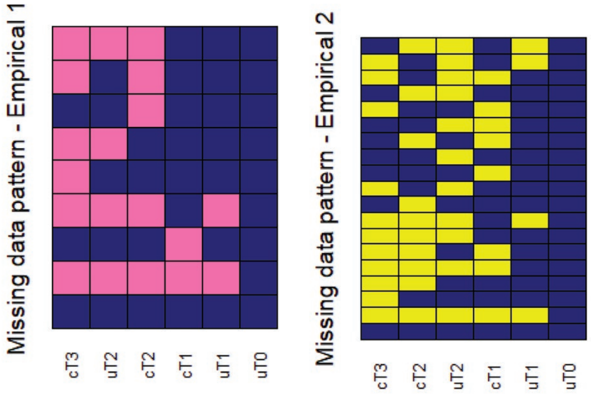


Supplementary Figure 1 | QALYs and total cost distributions for the control and the intervention groups in empirical dataset 1 and empirical dataset 2.

Supplementary Table 2 | Descriptive statistics of the empirical datasets

Control	n	uT0, mean (SD)	n	uT1 mean (SD)	n	uT2 mean (SD)	n	cT1 mean (SD)	n	cT2 mean (SD)	n	cT3 mean (SD)
Empirical data 1, = 169												
Complete	63	0.43 (0.29)	63	0.82 (0.21)	63	0.78 (0.25)	63	685 (1011)	63	790 (2741)	63	819 (1769)
Missings	14	0.29 (0.22)	10	0.70 (0.27)	8	0.73 (0.27)	9	1444 (1595)	8	2637 (4128)	5	1215 (1031)
Intervention												
Complete	78	0.41 (0.30)	78	0.81 (0.22)	78	0.78 (0.25)	78	834 (744)	78	495 (742)	78	515 (820)
Missings	14	0.35 (0.32)	6	0.83 (0.18)	4	0.92 (0.09)	5	1586 (1975)	4	20 (39)	3	79 (137)
Empirical data 2, n = 143												
Control	**n**	**uT0, mean (SD)**	**n**	**uT1 mean (SD)**	**n**	**uT2 mean (SD)**	**n**	**cT1 mean (SD)**	**n**	**cT2 mean (SD)**	**n**	**cT3 mean (SD)**
Complete	24	0.61 (0.29)	24	0.65 (0.32)	24	0.64 (0.22)	24	601 (764)	24	1162 (2590)	24	506 (684)
Missings	45	0.60 (0.32)	32	0.75 (0.22)	21	0.68 (0.25)	45	383 (409)	14	1511 (3932)	5	5515 (8749)
Intervention												
Complete	24	0.63 (0.29)	24	0.69 (0.27)	24	0.59 (0.24)	24	591 (1143)	24	381 (373)	24	1441 (4045)
Missings	50	0.61 (0.37)	38	0.66 (0.31)	21	0.65 (0.29)	50	418 (365)	12	274 (299)	8	2692 (6281)

uT0: utility at baseline. uT1: utility at time point 1. uT2: utility at time point 2. cT1: costs at time point 1. cT2: costs at time point 2. cT3: costs at time point 3. Costs were not collected at baseline in both randomized clinical trials.



Supplementary Figure 2 | Missing data pattern in empirical dataset 1 (pink) and empirical dataset 2 (yellow), respectively. Blue squares represent complete data. Coloured squares represent missing data. uT0: utility at baseline. uT1: utility at time point 1. uT2: utility at time point 2. cT1: costs at time point 1. cT2: costs at time point 2. cT3: costs at time point 3.

Supplementary Material 5

Joint Longitudinal Linear Mixed-model analysis – In this strategy, a joint LLM for costs and utility values was specified according to Faria et al. (2014), instead of two separated LLM models. That is, after rescaling costs to the same scale as utility values (i.e. 0-1), both outcomes were stacked on top of each other and simultaneously regressed upon the various covariates in the model using a three-level structure (i.e. subject, outcome, time) (please see Stata code in the next page). Results are presented in the table below.

Supplementary Table 3 | Performance of the joint longitudinal linear mixed-models for Costs and QALY

	Costs, €	JOINT-LLM	M-JOINT-LLM	MI-LLM	SUR-CCA	M-SUR	MI-SUR
Complete data	EB (MCse)	0.19 (3)	NA	NA	0.17 (3)	NA	NA
	RMSE (MCse)	153 (27)	NA	NA	153 (27)	NA	NA
	CR (MCse)	0.995 (0.158)	NA	NA	0.957 (0.456)	NA	NA
Missing 10%	EB (MCse)	-30 (4)	-92 (4)	-6 (3)	-106 (4)	-92 (4)	-6 (3)
	RMSE (MCse)	164 (29)	184 (32)	157 (27)	199 (34)	184 (32)	157 (27)
	CR (MCse)	0.939 (0.533)	0.898 (0.675)	0.948 (0.496)	0.913 (0.630)	0.895 (0.685)	0.948 (0.496)
Missing 25%	EB (MCse)	-41 (4)	-97 (4)	-10 (3)	-163 (6)	-97 (4)	-10 (3)
	RMSE (MCse)	182 (32)	194 (34)	157 (27)	269 (47)	195 (34)	157 (27)
	CR (MCse)	0.941 (0.527)	0.866 (0.760)	0.955 (0.463)	0.892 (0.694)	0.855 (0.787)	0.951 (0.483)
Missing 50%	EB (MCse)	-43 (5)	-224 (7)	-38 (5)	-167 (6)	-224 (7)	-38 (5)
	RMSE (MCse)	223 (40)	333 (56)	214 (38)	285 (49)	333 (56)	214 (38)
	CR (MCse)	0.905 (0.656)	0.583 (1.115)	0.933 (0.560)	0.860 (0.776)	0.507 (1.118)	0.925 (0.589)

	QALY	JOINT-LLM	M-JOINT-LLM	MI-LLM	SUR-CCA	M-SUR	MI-SUR
Complete data	EB	-0.0000589	NA	NA	-0.0000586	NA	NA
	(MCse)	(0.0001004)			(.0001004)		
	RMSE	.0044895	NA	NA	0.0044890	NA	NA
	(MCse)	(.0008193)			(0.0008194)		
	CR	0.914	NA	NA	0.948	NA	NA
	(MCse)	(0.625)			(0.496)		
Missing 10%	EB	-0.00000736	-0.004345	-0.000184	0.0010317	-0.0043359	-0.0001803
	(MCse)	(0.0001051)	(0.0001457)	(0.0001054)	(0.0001113)	(0.0001455)	(0.0001053)
	RMSE	0.0046984	0.0065157	0.0047126	0.0049750	0.0065048	0.0047078
	(MCse)	(0.0008556)	(0.0011049)	(0.0008573)	(0.0009084)	(0.0011014)	(0.0008568)
	CR	0.899	0.729	0.907	0.951	0.847	0.925
	(MCse)	(0.674)	(0.993)	(0.649)	(0.483)	(0.804)	(0.476)
Missing 25%	EB	-0.0000763	-0.0026896	-0.0000729	0.0016466	-0.0026772	-0.0000684
	(MCse)	(0.0001153)	(0.0001318)	(0.0001072)	(0.0001443)	(0.0001314)	(0.0001071)
	RMSE	0.0051572	0.0058925	0.0047946	0.0064497	0.0058766	0.0047889
	(MCse)	(0.0009329)	(0.0010421)	(0.0008518)	(0.0011593)	(0.0010395)	(0.0008516)
	CR	0.903	0.791	0.904	0.950	0.860	0.944
	(MCse)	(0.662)	(0.909)	(0.659)	(0.487)	(0.774)	(0.514)
Missing 50%	EB	-0.0001245	-0.02207	-0.0001778	0.00167082	-0.0220928	-0.00177395
	(MCse)	(0.0001424)	(0.0005329)	(0.0001463)	(0.0001542)	(0.0005334)	(0.0001462)
	RMSE	0.00636641	0.023827	0.0065429	0.0068928	0.0238469	0.00653833
	(MCse)	(0.0011428)	(0.0030349)	(0.0011727)	(0.0012258)	(0.0030351)	(0.0011710)
	CR	0.865	0.045	0.875	0.932	0.085	0.926
	(MCse)	(0.764)	(0.463)	(0.738)	(0.561)	(0.625)	(0.583)

JOINT-LLM: joint longitudinal linear mixed-model (i.e., costs and utilities at each time point were simultaneously regressed upon the various covariates in the model after rescaling their values). M-JOINT-LLM: Mean imputation combined with JOINT-LLM. MI-JOINT-LLM: Multiple imputation combined with JOINT-LLM. SUR-CCA: Seemingly unrelated regressions - complete case analysis. M-SUR: mean imputation combined with SUR. MI-SUR: multiple imputation combined with SUR. MCse: Monte Carlo standard error. EB: empirical bias. RMSE: root-mean-square error. CR: coverage rate. QALY: quality-adjusted life-year. €: Euros.

/* Stata code - JOINT LONGITUDINAL LINEAR MIXED-MODEL */

```
 clear
 set more off
 cd “C:\MISSINGX”
 local n = 2000

forvalues j = 1(1)`n’{
 local y = `j’
 use “dataset`y’”, clear

 //costs are scaled down by 7000 to transform them into a similar scale as utility values
 replace cT0 = cT0/7000
 replace cT1 = cT1/7000
 replace cT2 = cT2/7000
 replace cT3 = cT3/7000
 replace cT4 = cT4/7000

 //keep baseline values for costs and utilities in wide format
 gen utility_b = uT0
 gen cost_b = cT0

 //reshape from wide to long creating a new variable - time - that indicates time period
 reshape long cT uT, i(id) j(time 0 1 2 3 4)
 rename cT cost
 rename uT utility
 rename cost y1
 rename utility y2

 //reshape again to create a single dependent variable - y. The variable type indicates whether it refers to costs or utilities
 reshape long y, i(id time) j(type)
 gen cost=type==1
 gen QALY=type==2
 egen timetype=group(time type)

 //JOINT-LLM
  quietly mixed y i.cost#i.time i.cost#i.trt#i.time i.cost#i.time#c.utility_b i.cost#i.time#c.cost_b i.cost#i.time#c.age i.cost#i.time#i.gender || id:
 mat betaCE = e(b) /* extract matrix of betas */
 mat vari = e(V) /* extract matrix of variances */
 mat obs = e(N_g) /* extract number of observations used in the model */
 gen obs = obs[1,1]
 gen Za = 1.95996

 /*i.cost#i.year represents the interaction between the cost and utilities and each time point;
 i.cost#i.trt#i.time represents the effect of treatment (trt) on costs and utilities at each time point;
 i.cost#i.time#c.utility_b represents the effect of utility at baseline on costs and utilities at each time point.
 i.cost#i.time#c.cost_b represents the effect of cost at baseline on costs and utilities at each time point. */
```

```
//gen variables for utility difference at each time point adjusted for baseline costs
 gen utility_diff1 = betaCE[1,17]
 gen utility_diff2 = betaCE[1,18]
 gen utility_diff3 = betaCE[1,19]
 gen utility_diff4 = betaCE[1,20]

 //calculate QALY difference
  gen QALY_diff = (0.5*(utility_diff1+utility_diff1)*(3/12)) + (0.5*(utility_diff1+utility_diff2)*(3/12)) + (0.5*(utility_diff2+utility_diff3)*(3/12)) + (0.5*(utility_diff3+utility_diff4)*(3/12))

 //gen variables for variance utility difference at each time point
 gen varu1 = vari[17,17]
 gen varu2 = vari[18,18]
 gen varu3 = vari[19,19]
 gen varu4 = vari[20,20]

 //calculate variance of QALY difference
 gen QALY_var = ((0.25*(varu1+varu1)) + (0.25*(varu1+varu2)) + (0.25*(varu2+varu3)) + (0.25*(varu3+varu4)))/4

 //calculate SE of QALY difference
 gen SE_QALY_diff = sqrt(QALY_var)

 //estimate CI around QALY difference
 gen LL_QALY = QALY_diff - Za*SE_QALY_diff
 gen UL_QALY = QALY_diff + Za*SE_QALY_diff

 //gen variables for cost difference at each time point adjusted for baseline costs
 gen cost_diff1 = betaCE[1,27]
 gen cost_diff2 = betaCE[1,28]
 gen cost_diff3 = betaCE[1,29]
 gen cost_diff4 = betaCE[1,30]

 //calculate marginal cost difference
 gen cost_diff = (cost_diff1 + cost_diff2 + cost_diff3 + cost_diff4)*7000

 //gen variables for variance of cost difference at each time point
 gen varc1 = vari[27,27]
 gen varc2 = vari[28,28]
 gen varc3 = vari[29,29]
 gen varc4 = vari[30,30]

//calculate SE for marginal cost difference
 gen SEc1 = sqrt(varc1)
 gen SEc2 = sqrt(varc2)
 gen SEc3 = sqrt(varc3)
 gen SEc4 = sqrt(varc4)
 gen SE_cost_diff = (SEc1 + SEc2 + SEc3 + SEc4)*7000

 //estimate CI around marginal cost difference
 gen LL_costs = (cost_diff - Za*SE_cost_diff)
 gen UL_costs = (cost_diff + Za*SE_cost_diff)

 //calculate covariance costs and QALY
 gen cov = vari[2,17]*7000

 save “JOINT-LLM\postboots`y’”, replace
}
```

/* MULTIPLE IMPUTATION + JOINT LONGITUDINAL LINEAR MIXED-MODEL*/

```
 clear
 set more off
 cd “C:\MISSINGX”
 local n = 2000

forvalues j = 1(1)`n’{
 local y = `j’
 use “dataset`y’”, clear

// MULTIPLE IMPUTATION MODEL
 mi set flong
 mi register regular age gender trt uT0 cT0
 mi register imputed cT1 cT2 cT3 cT4 uT1 uT2 uT3 uT4
 quietly mi impute chained (pmm, knn(5)) cT1 cT2 cT3 cT4 uT1 uT2 uT3 uT4 = age gender uT0 cT0, by(trt) replace
add(10) rseed(`k’)
 save «C:\MISSINGX\MI-JOINT-LLM\dataset`y’_imp», replace
}

 clear
 set more off
 cd «C:\MISSINGX\MI-JOINT-LLM»
 local n = 2000

forvalues k=1(1)`n’{
 local y = `k’
 use «dataset`y’_imp», clear

//costs are scaled down by 7000 to transform them into a similar scale as utilities
 replace cT0 = cT0/7000
 replace cT1 = cT1/7000
 replace cT2 = cT2/7000
 replace cT3 = cT3/7000
 replace cT4 = cT4/7000

//keep baseline values for costs and utilities in wide format
 gen utility_b = uT0
 gen cost_b = cT0

//reshape from wide to long creating a new variable - time - that indicates time period
 rename (cT0 cT1 cT2 cT3 cT4) (cost0 cost1 cost2 cost3 cost4)
 rename (uT0 uT1 uT2 uT3 uT4) (utility0 utility1 utility2 utility3 utility4)
 quietly mi reshape long cost utility, i(id) j(time 0 1 2 3 4)
 rename cost y1
 rename utility y2

//reshape again to create a single dependent variable - y. The variable type indicates whether it refers to costs or
utilities
 mi reshape long y, i(id time) j(type)
 gen cost=type==1
 gen QALY=type==2
 egen timetype=group(time type)

//JOINT-LLM performed in each one of the imputed datasets (MI-JOINT-LLM)
  quietly mi estimate: mixed y i.cost#i.time i.cost#i.trt#i.time i.cost#i.time#c.utility_b i.cost#i.time#c.cost_b i.cost#i.
time#c.age i.cost#i.time#i.gender|| id:
mat betaCE = e(b_mi) /* extract matric of betas */
 mat vari = e(V_mi) /* extract matrix of variances */
 mat obs = e(N_g_mi) /* extract number of observations used in the model */
 gen obs = obs[1,1]
 gen loss_eff = e(fmi_max_mi)/e(M_mi) /* calculate loss of efficiency)
 gen Za = 1.95996

//gen variables for variance utility difference at each time point
 gen utility_diff1 = betaCE[1,17]
 gen utility_diff2 = betaCE[1,18]
 gen utility_diff3 = betaCE[1,19]
 gen utility_diff4 = betaCE[1,20]



//calculate QALY difference
gen QALY_diff = (0.5*(utility_diff1+utility_diff1)*(3/12)) + (0.5*(utility_diff1+utility_diff2)*(3/12)) + (0.5*(utility_
diff2+utility_diff3)*(3/12)) + (0.5*(utility_diff3+utility_diff4)*(3/12))

//calculate variance of QALY difference
 gen varu1 = vari[17,17]
 gen varu2 = vari[18,18]
 gen varu3 = vari[19,19]
 gen varu4 = vari[20,20]

//calculate variance of QALY difference
 gen QALY_var = ((0.25*(varu1+varu1)) + (0.25*(varu1+varu2)) + (0.25*(varu2+varu3)) + (0.25*(varu3+varu4)))/4

 //calculate SE of QALY difference
 gen SE_QALY_diff = sqrt(QALY_var)

 //estimate CI around QALY difference
 gen LL_QALY = QALY_diff - Za*SE_QALY_diff
 gen UL_QALY = QALY_diff + Za*SE_QALY_diff

//gen variables for cost difference at each time point adjusted for baseline costs
 gen cost_diff1 = betaCE[1,27]
 gen cost_diff2 = betaCE[1,28]
 gen cost_diff3 = betaCE[1,29]
 gen cost_diff4 = betaCE[1,30]

//gen variables for variance of cost difference at each time point
 gen cost_diff = (cost_diff1 + cost_diff2 + cost_diff3 + cost_diff4)*7000

//gen variables for variance of cost difference at each time point
 gen varc1 = vari[27,27]
 gen varc2 = vari[28,28]
 gen varc3 = vari[29,29]
 gen varc4 = vari[30,30]

//calculate SE for marginal cost difference
 gen SEc1 = sqrt(varc1)
 gen SEc2 = sqrt(varc2)
 gen SEc3 = sqrt(varc3)
 gen SEc4 = sqrt(varc4)
 gen SE_cost_diff = (SEc1 + SEc2 + SEc3 + SEc4)*7000

//estimate CI around marginal cost difference
 gen LL_costs = (cost_diff - Za*SE_cost_diff)
 gen UL_costs = (cost_diff + Za*SE_cost_diff)

//calculate covariance costs and QALY
 gen cov = vari[2,17]*7000
 save «postboots`y'», replace
}
```

/* MEAN IMPUTATION + JOINT LONGITUDINAL LINEAR MIXED-MODEL */

```
 clear
 set more off
 cd «C:\MISSING10»
 local n = 2000

forvalues k=1(1)`n'{
 local y = `k'
 use «dataset`y'», clear

// MEAN IMPUTATION BY TREATMENT GROUP
 bysort trt: egen mean_uT1 = mean(uT1)
 bysort trt: egen mean_uT2 = mean(uT2)
 bysort trt: egen mean_uT3 = mean(uT3)
 bysort trt: egen mean_uT4 = mean(uT4)

 replace uT1 = mean_uT1 if missing(uT1)
 replace uT2 = mean_uT2 if missing(uT2)
 replace uT3 = mean_uT3 if missing(uT3)
 replace uT4 = mean_uT4 if missing(uT4)

 bysort trt: egen mean_cT1 = mean(cT1)
 bysort trt: egen mean_cT2 = mean(cT2)
 bysort trt: egen mean_cT3 = mean(cT3)
 bysort trt: egen mean_cT4 = mean(cT4)

 replace cT1 = mean_cT1 if missing(cT1)
 replace cT2 = mean_cT2 if missing(cT2)
 replace cT3 = mean_cT3 if missing(cT3)
 replace cT4 = mean_cT4 if missing(cT4)

//costs are scaled down by 7000 to transform them into a similar scale as utilities
 replace cT0 = cT0/7000
 replace cT1 = cT1/7000
 replace cT2 = cT2/7000
 replace cT3 = cT3/7000
 replace cT4 = cT4/7000

 //keep baseline values for costs and utilities in wide format
 gen utility_b = uT0
 gen cost_b = cT0

 //reshape from wide to long creating a new variable - time - that indicates time period
 reshape long cT uT, i(id) j(time 0 1 2 3 4)
 rename cT cost
 rename uT utility
 rename cost y1
 rename utility y2

 //reshape again to create a single dependent variable - y. The variable type indicates whether it refers to costs or
utilities
 reshape long y, i(id time) j(type)
 gen cost=type==1
 gen QALY=type==2
 egen timetype=group(time type)

 //JOINT-LLM performed in the mean imputed data (M-JOINT-LLM)
  quietly mixed y i.cost#i.time i.cost#i.trt#i.time i.cost#i.time#c.utility_b i.cost#i.time#c.cost_b i.cost#i.time#c.age
i.cost#i.time#i.gender || id:
 mat betaCE = e(b) /* extract matrix of betas */
 mat vari = e(V) /* extract matrix of variances */
 mat obs = e(N_g) /* extract number of observations used in the model */
 gen obs = obs[1,1]
 gen Za = 1.95996

//gen variables for utility difference at each time point adjusted for baseline costs
 gen utility_diff1 = betaCE[1,17]
 gen utility_diff2 = betaCE[1,18]
 gen utility_diff3 = betaCE[1,19]
 gen utility_diff4 = betaCE[1,20]
```

```
//calculate QALY difference
 gen QALY_diff = (0.5*(utility_diff1+utility_diff1)*(3/12)) + (0.5*(utility_diff1+utility_diff2)*(3/12)) + (0.5*(utility_diff2+utility_diff3)*(3/12)) + (0.5*(utility_diff3+utility_diff4)*(3/12))

 //gen variables for variance utility difference at each time point
 gen varu1 = vari[17,17]
 gen varu2 = vari[18,18]
 gen varu3 = vari[19,19]
 gen varu4 = vari[20,20]

 //calculate variance of QALY difference
 gen QALY_var = ((0.25*(varu1+varu1)) + (0.25*(varu1+varu2)) + (0.25*(varu2+varu3)) + (0.25*(varu3+varu4)))/4

 //calculate SE of QALY difference
 gen SE_QALY_diff = sqrt(QALY_var)

 //estimate CI around QALY difference
 gen LL_QALY = QALY_diff - Za*SE_QALY_diff
 gen UL_QALY = QALY_diff + Za*SE_QALY_diff

//gen variables for cost difference at each time point adjusted for baseline costs
 gen cost_diff1 = betaCE[1,27]
 gen cost_diff2 = betaCE[1,28]
 gen cost_diff3 = betaCE[1,29]
 gen cost_diff4 = betaCE[1,30]

 //calculate marginal cost difference
 gen cost_diff = (cost_diff1 + cost_diff2 + cost_diff3 + cost_diff4)*7000

 //gen variables for variance of cost difference at each time point
 gen varc1 = vari[27,27]
 gen varc2 = vari[28,28]
 gen varc3 = vari[29,29]
 gen varc4 = vari[30,30]

//calculate SE for marginal cost difference
 gen SEc1 = sqrt(varc1)
 gen SEc2 = sqrt(varc2)
 gen SEc3 = sqrt(varc3)
 gen SEc4 = sqrt(varc4)
 gen SE_cost_diff = (SEc1 + SEc2 + SEc3 + SEc4)*7000

 //estimate CI around marginal cost difference
 gen LL_costs = (cost_diff - Za*SE_cost_diff)
 gen UL_costs = (cost_diff + Za*SE_cost_diff)

 //calculate covariance costs and QALY
 gen cov = vari[2,17]*7000

 save “M-JOINT-LLM\postboots`y’”, replace
```

Part 2

Tutorials

CHAPTER 8

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

under embargo

CHAPTER 9

Interpretation of trial-based economic evaluations of musculoskeletal physical therapy interventions

Gisela Cristiane Miyamoto, Ângela Jornada Ben, Judith E. Bosmans, Maurits W. van Tulder, Christine C. Lin, Cristina Maria Nunes Cabral, Johanna M. van Dongen

Brazilian Journal of Physical Therapy, 2021; 25(5): 514 – 529

Abstract

Background

As resources for healthcare are scarce, decision-makers increasingly rely on economic evaluations when making reimbursement decisions about new health technologies, such as drugs, procedures, devices, and equipment. Economic evaluations compare the costs and effects of two or more interventions. Musculoskeletal disorders have a high prevalence and result in high levels of disability and high costs worldwide. Because physical therapy interventions are usually the first line of treatment for musculoskeletal disorders, economic evaluations of such interventions are becoming increasingly important for stakeholders in the field of physical therapy, including physical therapists, decision-makers, and researchers. However, economic evaluations are relatively difficult to interpret for the majority of stakeholders.

Methods

The design, analysis, and interpretation of economic evaluations performed alongside randomized controlled trials are discussed. To further illustrate and explain these concepts, we use a case study assessing the cost-effectiveness of exercise therapy compared to standard advice in patients with musculoskeletal disorders.

Conclusions

Economic evaluations are increasingly being used in healthcare decision-making. Therefore, it is of utmost importance that their design, conduct, and analysis are state-of-the-art and that their interpretation is adequate. This masterclass will help physical therapists, decision-makers, and researchers in the field of physical therapy to critically appraise the quality and results of trial-based economic evaluations and to apply the results of such studies to their own clinical practice and setting.

Introduction

Resources available for healthcare are scarce worldwide. Decision-makers increasingly request information on the relative efficiency of healthcare interventions when making reimbursement decisions. This information is provided by economic evaluations, which compare both the costs and effects of two or more interventions.[1] Although the use of economic evaluation results in healthcare decision-making is most common in high-income countries, low- and middle-income countries have recently acknowledged the importance of using such economic evidence in their healthcare decision-making process.[2,3,4,5]

In recent years, the prevalence of musculoskeletal disorders has increased exponentially, resulting in high levels of disability and high costs.[6,7,8,9,10] Musculoskeletal disorders are the leading cause of years lived with disability and work absence.[6,11] Low back pain and neck pain presented the highest healthcare costs (USD134.5 billion) in the United States between 1996-2016.[10] The total annual costs of low back pain alone are estimated at about USD15 billion in the United Kingdom and USD11 billion in Australia.[12,13] The Brazilian public healthcare system was found to spend approximately USD714 million on spinal disorders, and the societal costs of low back pain alone between 2012-2016 were about USD2.2 billion.[14,15] However, some of the healthcare budget for low back pain is spent on unnecessary diagnostic tests or on not recommended interventions.[14,16,17] Economic evaluations may help healthcare decision-makers on how to allocate these scarce resources as efficiently as possible.[18] Consequently, low-, middle-, and high-income countries have started using economic evaluation as an input for reimbursement decisions.[19,20,21,22]

Physical therapy interventions are the first line of treatment for many musculoskeletal disorders.[23,24,25] Economic evaluations of physical therapy interventions are becoming increasingly important for stakeholders in the field of physical therapy, including researchers, physical therapists, and decision-makers.[2,26] However, the uptake of economic evaluation results among those stakeholders is hampered by the fact that for many of them economic evaluations are generally complex and difficult to interpret. Evidence indicates, for example, that although healthcare decision-makers are highly interested in economic evaluations, the impact of the results of such studies has been limited due to a lack of knowledge and skills required to interpret the results.[19,27,28,29,30,31] This masterclass aimed to support stakeholders commissioned with making decisions about the treatment of musculoskeletal disorders with the interpretation of economic evaluations and translating the results of such studies into clinical practice.

International recommendations

This masterclass is based on the most recent international recommendations for trial-based economic evaluations.[18,22,32,33,34,35,36] A case study consisting of an economic evaluation of exercise therapy for non-specific chronic low back pain in Brazil, is used as an example to illustrate how general principles regarding the design, analysis, and reporting of trial-based economic evaluations of musculoskeletal physical therapy interventions apply in such a specific setting. Information about the case study is presented in Box 1.[37,38]

Box 1 | Discounting of costs

The case study concerns an economic evaluation performed alongside a randomized controlled trial assessing the cost-effectiveness of an exercise therapy consisting of Pilates exercises compared to standard advice.[37] Two-hundred ninety-six patients were randomly allocated to four treatment groups: booklet, Pilates 1, Pilates 2, and Pilates 3. All patients received physical therapy advice. The booklet group did not receive other treatment recommendations. Pilates groups 1, 2, and 3 received individualized exercise therapy that were given once, twice, and three times a week, respectively, for six weeks. In this masterclass, we only used data from Pilates 3 group (exercise therapy group) and the booklet group (control group).[37] Patients with chronic non-specific low back pain, aged between 18 and 80 years were included. Patients with contraindications for exercise, pregnancy, nerve root compression, or serious spinal pathologies, and previous or scheduled spinal surgery were excluded. The study was conducted at a Pilates clinic and a physical therapy clinic. The effect outcomes were defined according to the core outcome set for low back pain.[38]

What is an economic evaluation?

Economic evaluations are defined as *"the comparative analysis of alternative courses of action in terms of both their costs and consequences"*.[1] Full economic evaluations identify, measure, value, and compare costs and health effects between two or more interventions.[1,18,33,36,39] Studies that do not compare costs and effects of two or more interventions are not considered full economic evaluations, but partial evaluations. Examples of such partial evaluations are cost-of-illness studies, in which only costs are considered, or cost-outcome descriptions, which only describe the costs and effects of one intervention.[18,33]

Design of an economic evaluation

Economic evaluations can be performed using decision analytical modelling techniques (i.e. model-based economic evaluations) or alongside randomized controlled trials (i.e. trial-based economic evaluations).[34,40] These two designs of economic evaluations are seen as complementary.[41]

In model-based economic evaluations, cost and effect data are obtained from different sources, such as systematic reviews, randomized controlled trials, cohort studies, electronic medical records, and other databases in which data are collected in daily practice.[22,40,42,43,44] These cost and effect estimates are then used as parameters in decision analytical models, such as decision trees, Markov models, and micro-simulations.[22,40,42,43,44] Model-based economic evaluations are useful when it is not possible to compare all relevant interventions in a trial, when trials do not assess all relevant costs and effects, or when decision-makers are interested in the long-term cost-effectiveness of interventions, while long-term individual patient-level data are lacking or impossible to collect prospectively.[22,40,42,43,44] To increase the quality of model-based economic evaluations, national and international guidelines for good practice have been published.[3,22,40,42,43,44]

In economic evaluations conducted alongside randomized controlled trials, patients are randomly allocated to one of the interventions and patient-level cost and effect data are gathered prospectively during follow-up.[32,34,36,41] It is also possible to conduct a trial-based economic evaluation alongside a non-randomized trial, such as a pre-post study. However, randomized controlled trials are generally considered the gold standard, because randomization ensures that all observed and unobserved confounders are equally distributed across groups, which improves the internal validity of the results. A possible disadvantage of randomized controlled trials is that their external validity (i.e., generalizability) is limited due to the selection of a restricted patient population and/or strict protocol for interventions. This can be improved by using a pragmatic trial design, which means that the trial is conducted under "real-world" conditions (i.e. resembling normal daily clinical practice).[32] Such a pragmatic design is considered the best study design for making inferences about the cost-effectiveness of healthcare interventions in clinical practice.[32,34,45,46] Therefore, this masterclass article focuses on trial-based economic evaluations.

Perspective

The perspective of an economic evaluation determines which costs and effects are included.[1,18,33,36,39] The broadest perspective is the societal perspective, in which all costs and effects are included, irrespective of who pays or benefits. The healthcare perspective is narrower and means that only costs borne by the healthcare sector are included. Other perspectives might also be relevant, such as that of the healthcare provider, the health insurance company, the patient, or the employer.[1,18,33,36,39] Because the applied perspective determines which cost categories are assessed and included in an economic evaluation, it should always be stated explicitly.[1,18,33,36,39]

Differences exist between countries regarding the recommended perspective. In the United Kingdom, for example, the healthcare perspective is recommended;[3] in Brazil, the Brazilian public healthcare system (*Sistema Único de Saúde* [SUS]) perspective;[22] and in the Netherlands the societal perspective.[20] An important advantage of the societal perspective is that it provides an estimation of the impact of implementing an intervention across all stakeholders.[36] This information will ensure that there is a net societal benefit (or loss), rather than simply costs shifting from one stakeholder to another. Moreover, a disaggregated presentation of the societal costs provides a good indication of their distribution among stakeholders.[36] All relevant cost categories are included in the societal perspective, making it possible to easily conduct additional analyses from a narrower perspective.[36,47] In the case study, the economic evaluation was conducted from the societal perspective and an additional analysis was performed from the narrower SUS perspective.[37]

Time horizon

The time horizon of an economic evaluation is the period over which cost and effect data are collected and analyzed.[1,36,39] This period should be long enough to allow for the assessment of all relevant costs and effects flowing from the intervention under study.[1,36,39] The most appropriate time horizon also depends on the nature of the health problem (e.g. acute, sub-acute, or chronic), the duration of the intervention under study, and the expected retention of the effect(s) of the intervention.[1,22,36,39] For example, a 12-week follow-up might be long enough to assess the cost-effectiveness of paracetamol and diclofenac compared with advice alone for *acute* low back pain.48 However, in the case study patients were suffering from *chronic* non-specific low back pain and Pilates-based exercise therapy was expected to improve pain and disability, as well as improve motor control, stabilization, and body awareness in the long-term.[37] Therefore, a 12-month time horizon was used. In general, researchers and physical therapists should at least feel confident that the most important costs and effects are covered by the chosen time horizon.[36] Even though the optimal follow-up period of trial-based economic evaluations of musculoskeletal intervention is unknown,[36] most studies in this area use a follow-up period of at least 12 months.[49,50,51,52]

Identification, measurement, and valuation of effects

The effect outcome that is measured and included in the analysis of an economic evaluation determines the type of economic evaluation.[1,18,33,34,36,39]

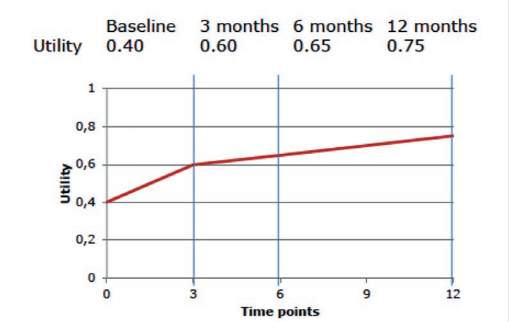
In a cost-effectiveness analysis (CEA), effect outcomes are disease- and/or intervention-specific and are particularly relevant for healthcare providers who use these measures to make decisions about the treatment of their patients.[53] In the majority of cases, this outcome is the primary outcome of a randomized controlled trial. In the field of musculoskeletal disorders, several core outcomes sets have been developed.[38,54,55,56] In patients with low back pain, for example, it is recommended to measure physical functioning using the Oswestry Disability Index or the Roland-Morris Disability Questionnaire, pain intensity using a numerical rating scale, and health-related quality of life (QoL) using the Short-Form 12 or PROMIS Global Health.[38] Such disease-specific outcomes are also recommended for other musculoskeletal disorders.[54,55,56] Because disease-specific outcomes are specific for the health condition and intervention under study, it is only possible to compare results of CEA across different types of musculoskeletal disorders when the same clinical outcome (e.g. pain intensity) was assessed. However, when decision-makers need to choose between reimbursing a treatment for musculoskeletal disorders or other conditions, such as cancer and diabetes, CEAs are of little use.[1,18,33,36,39]

In a cost-utility analysis (CUA), effects are measured in terms of Disability-Adjusted Life-Years (DALYs) or Quality-Adjusted Life-Years (QALYs).[1,18,33,34,36,39,57] Both DALYs and QALYs combine morbidity and life-expectancy in one single measure. This allows for making comparisons across different kinds of health conditions and interventions.[57] Whereas QALYs represent the life-years spent in optimal health, DALYs represent the loss of quality of life due to health conditions.[58] Although the World Health Organization recommends the use of DALYs for economic evaluations, most national pharmacoeconomic organizations, such as the National Institute for Health and Clinical Excellence (NICE), the Dutch National Health Care Institute, and the Rede Brasileira de Avaliação de Tecnologias em Saúde (REBRATS) recommend to use QALYs for the purpose of healthcare decision-making.[20,21,22] To estimate QALYs, three steps are typically followed in trial-based economic evaluations: 1) assessment of the patients' health states using a preference-based QoL measure, 2) conversion of the patients' health states into utility values, and 3) calculation of QALYs by multiplying the patients' utility values by the time they spent in a specific health state.

Preference-based QoL measures that can be used are the EuroQol 5 Dimensions (EQ-5D), the Health Utilities Index (HUI), and the Short-Form 6 Dimensions (SF-6D, which can be derived from the SF-12 and SF-36 questionnaires).[59,60,61] These questionnaires are ideally administered at different time points to describe the participants' QoL during the course of the trial. The more often QoL is measured, the more precise the estimate of effect, although frequent assessment may be burdensome to patients.[62] To convert the patients' health states to utility values, national value sets are typically used, in which each health state is converted to a utility value previously derived from the preferences of the general population.[62] Utility values indicate a person's preference for a specific health state on a scale thats is anchored at 0 (equal to death) and 1 (equal to full health).[62] Negative values can also occur and indicate that a specific health state is considered to be worse than death.[62] Finally, the obtained utility values are used to calculate QALYs by multiplying them by the amount of time a patient spent in a specific health state.[34] An example of such a calculation is presented in Box 2.

Box 2 | Estimating QALYs

1) Assessment of the patients' health states using a preference-based QoL measure;
2) Conversion of the patients' health states into utility values;
3) Calculation of QALYs using linear interpolation between measurement points



To calculate the QALYs using a hypothetical participant's QoL at baseline (utility value: 0.4), and at 3-month (utility value: 0.6), 6-month (utility value: 0.65), and 12-month (utility value: 0.75) follow-up, we first need to estimate the average utility value per measurement period.
For the first period (baseline to 3 months), this is (0.4 + 0.6) / 2 = 0.5.
For the second period (3 months to 6 months), this is (0.6 + 0.65) / 2 = 0.625, and for the third period (6 months to 12 months), this is (0.65 + 0.75) / 2 = 0.70. Subsequently, we need to multiply these average utility values per time period by the length of that time period, i.e., the time spent in a particular health state, and sum them all up. Thus, this participant's number of QALYs gained during the 12-month follow-up period is calculated as follows:
QALY=((0.4+0.6/2)*(3/12))+((0.6+0.65/2)*(3/12))+((0.65+0.75/2)*(6/12))=0.62
QALY can range from 0 to 1, where 0 indicates "death" and 1 indicates "full health".

In the case study, the patients' health states were measured using the SF-6D and converted to utility values using the Brazilian tariff.[61,63] QALYs were calculated using linear interpolation between measurement points (Box 2).

There are two other types of economic evaluations, which are not frequently used in health research. In a cost-benefit analysis (CBA), both costs and effects are expressed in monetary units. CBAs provide an indication of whether an intervention generates savings or losses compared with an alternative and are also referred to as return-on-investment analyses.[1,18,33,36,39] However, monetizing clinical outcomes, such as pain, disability, and recovery is considered difficult and even unethical sometimes. Therefore, these analyses are considered less relevant in the evaluation of physical therapy interventions.[36]

Finally, in a cost-minimization analysis (CMA), effects are considered equal for the interventions compared, and therefore only costs are compared between the alternatives.[1,18,33,36,39] This approach, however, does not take into account the joint uncertainty surrounding the costs and effects of interventions. Also, a conclusion that effects are equal can only be made if the study was designed specifically to demonstrate equivalence of the compared interventions. Absence of a statistically significant difference cannot be considered evidence of equivalence.[64] Unless a study sets out to show equivalence of two treatments, CMAs are considered inappropriate.[1,18,33,36,39,64]

In practice, most economic evaluations are a combination of a CEA (to inform healthcare providers) and a CUA (to inform healthcare decision-makers).[18,33] In the case study, two CEAs were performed, i.e. one for physical functioning and one for pain intensity, and a CUA was performed for QALYs.[37]

Identification, measurement, and valuation of costs

An integral part of any economic evaluation is the identification, measurement, and valuation of the resources consumed by the patients.[1,18,34,36,39] Resources are, for example, number of pills taken, number of visits to a general practitioner or a physical therapist, or the performance of a diagnostic test. The resource use items that need to be included highly depend on the applied perspective, the interventions being evaluated, and the patient population.[1,18,34,36,39] Once all relevant resource categories are identified, researchers should determine how to "cost" them. This process involves three steps: 1) the measurement of the quantities of resources consumed (Q), 2) the assignment of unit prices (p), and 3) the valuation of the resources consumed ($C = Q * p$).[1,65] Ideally, the quantities of resources consumed as well as their respective unit prices are reported separately so that readers can recalculate costs for their own setting.[1,65] The "costing" steps will be discussed below into more detail.

The measurement of quantities of resources consumed (Q)

Resource use data can be collected using patient medical records, insurance records, interviews, questionnaires, cost diaries, previous studies, information from vendors, and/or administrative databases.[1,18,32,33,36,66,67] Several questionnaires for assessing resource use have been developed (e.g. iMTA Questionnaire on Costs, iMTA Valuation of Informal Care Questionnaire).[68,69] Researchers typically develop their own cost questionnaire, based on existing questionnaires, to tailor it to their specific population, and are encouraged to publish these questionnaires in an open-access database (e.g. Database of Instruments for Resource Use Measurement [DIRIUM]).[70]

If medical or insurance records are used, recall bias (i.e., risk of patients forgetting information) is non-existent. However, such databases may lack important information (i.e., information bias) because it is simply not recorded or measured, for example information on healthcare utilization that is not reimbursed by the insurer. It is also possible that information is collected incompletely, because reimbursement is based on a package of care (e.g. diagnosis related groups, multidisciplinary treatments) and not on the separate resource utilization items.[18] Although this may not be problematic when using the healthcare insurer perspective, this is not appropriate when using the societal perspective.

If patient self-reports are used, a balance needs to be found between the duration of the recall periods and the frequency with which the instrument is administered. This is important because the risk of recall bias increases with longer recall periods, whereas increasing the number of assessments increases the burden for the participants. When relatively short recall periods (e.g., only a couple of weeks) are used over a longer period of time, this may be overly burdensome to patients, which may increase the risk of missing data and drop-outs. To minimize recall bias, missing data, and drop-outs, the literature recommends recall periods of two to six months in a study with a long-term follow-up (e.g. more than 12 months).[34,71,72] It might be useful to measure healthcare utilization more frequently during the first months of a physical therapy study, because most healthcare utilization and most sick leave will occur when patients are seeking healthcare for a new episode of musculoskeletal complaints. In the case study, patients were asked to fill in a cost diary assessing all resources used related to their low back pain symptoms. This information was collected by telephone every six weeks during a period of 12 months.[37]

The assignment of unit prices (p)

Ideally, unit prices reflect opportunity costs, which are defined as *"the value of a resource in its most highly valued alternative use"*.[1,73] In simple terms, opportunity costs are equal to not receiving the benefit of the next best option. As such, opportunity costs are thought to reflect the value of the actual resources used. Charges or tariffs do not reflect opportunity costs or the actual value, because they are based on negotiations, e.g. between the government and healthcare organizations.[33,34] Therefore, they should not be used in economic evaluations. Unit price information can be obtained from national databases (e.g. SUS cost table), costing manuals (e.g. Dutch manual), professional organizations, previous studies, vendors, and/or administrative databases.[18,20,74,75,76,77]

The valuation of resources consumed (C=Q*p)

Valuation is the process of converting resource utilization rates into costs by multiplying them with their opportunity costs. Resources (C) are valued by multiplying the quantities of resources consumed (Q) with the unit prices (p) (C = Q*p). Below, a more detailed description of the identification, measurement, and valuation of resources is provided for cost categories that are often included in trial-based economic evaluations of musculoskeletal physical therapy interventions.

Intervention costs

If the cost of an intervention is unknown, it can be estimated using a micro-costing or a gross-costing approach.[1,33,36,78,79] In a micro-costing approach (i.e. bottom-up approach), information on the types and quantities of resources consumed as well as their respective unit prices is collected for each intervention component separately.[79] In the case study, for example, the components of the exercise therapy included Pilates exercise sessions and education materials.[37] For each of those components, information was gathered about the staff involved as well as the number of hours that they devoted to providing the intervention, the materials used, the housing needed, and the associated overhead costs (e.g. cleaning costs, costs of heating).[1,33,36] The quantities of resources consumed per intervention component can be measured through interviews or surveys with providers and/or patients, expert panels, administrative databases, intervention logs, or observations.[36,80] Micro-costing gives a reliable and precise estimate of the intervention costs, but is time-consuming. A gross-costing approach is simpler, and therefore less time-intensive. It allocates a total budget to specific services, such as physical therapists' visits, using specific allocation rules.[1,33,36,78,81,82] The average intervention cost per patient might, for example, be estimated by simply dividing the total intervention costs by the number of patients. Although gross-costing is a simple and fast approach, it lacks precision, and its success depends on the type of routine data available. Thus, the choice between micro-costing and gross-costing depends on how large the contribution of a specific cost item is to the total costs. Many studies use a mix of both approaches, for example, by using micro-costing for estimating intervention costs and gross-costing for all other cost categories.[78,81, 82, 83]

Healthcare utilization costs

Ideally, the use of all healthcare services is measured to reduce the likelihood of missing important, but unexpected shifts in healthcare services use.[84] Although this approach increases the validity of the results, it might not always be feasible.[84] An alternative strategy is to limit data collection to healthcare utilization that is deemed to be related to the health condition under study and those expected to differ between the interventions.[84] Healthcare utilization generally includes, amongst others, the use of medications, primary care services (e.g. number of visits to general practitioners or other healthcare professionals, physical therapy sessions, diagnostic tests), secondary care services (e.g. number of outpatient hospital visits, visits to other healthcare institutions such as a rehabilitation clinic, and admissions to hospital), and tertiary care services (e.g. number of visits to a specialized clinic with highly specialized medical care).[1,18,33,36]

Patient and family costs

Patient and family costs include all costs accruing to patients and/or their family members, including costs of over-the-counter medications and transportation, but also informal care costs.[1,18,33,36] Informal care refers to paid and unpaid activities by one or more members of the social environment of the patient.[85] Informal care tasks may comprise housekeeping, personal care, support with mobility, and administrative tasks.[85] In economic evaluations of physical therapy interventions, informal care can be an important cost category, because an increasing part of the total care provided to patients, especially to patients with chronic diseases, consists of informal care.[85] Failure to include this category will result in an underestimation of total societal costs, and possibly to missing important shifts from formal care to informal care. Different approaches can be used to value informal care. The most widely used option is the use of a shadow price, e.g., the hourly costs of a legally employed cleaner. Other approaches are the proxy good approach, where the costs of a market substitute (e.g. the hourly wage rate of a nurse for nursing tasks, the hourly costs of a legally employed cleaner for cleaning tasks) are used, and the opportunity cost approach, where the actual wage rate of the informal care giver is used.[85]

Lost productivity costs

Productivity losses are an important cost driver in many economic evaluations of physical therapy interventions. Musculoskeletal disorders often lead to reduced productivity, because patients cannot perform their work and therefore report in sick or become less productive at work.[11] Productivity loss is defined as a loss of labour output (e.g. a company's output) as a result of reduced labour input (i.e. time and efforts of workers with a health problem).[86] Thus, productivity loss is ideally estimated by measuring output loss. However, it is difficult to estimate the true impact of the reduced labour input on a company's output.[36,87,88] Therefore, researchers typically use proxies of productivity loss, which include losses related to reduced productivity while at work (i.e. presenteeism) and losses related to absence from paid work (i.e. absenteeism) using self-reported data.[1,18,33,36]

Research indicates that presenteeism often represents a large part of total productivity losses.[86] Several questionnaires are available for assessing presenteeism, including the World Health Organization Health and Work Performance Questionnaire, the Quantity and Quality questionnaire, the Work Limitations Questionnaire, and the iMTA Productivity Cost Questionnaire.[89,90,91,92,93,94,95,96] These questionnaires typically ask patients to rate their work performance in terms of points, percentages, or a proportion compared to their normal performance. These outcomes can then be used to estimate the number of days lost due to presenteeism using the following fomula:[97]

$$\textit{Presenteeism days} = (T - S) * (1 - w)$$

where *Presenteeism days* is the number of days lost due to presenteeism, *T* is total number of working days, *S* is the total number of sickness absence days, and *w* is the patient's self-reported work performance.[86]

Absenteeism from paid work represents another important source of lost productivity, and, thus, societal costs.[86,87] There are two methods for valuing absenteeism from paid work, namely the Human Capital Approach and the Friction Cost Approach.[1,18,33,36,67] According to the Human Capital Approach, absenteeism costs are equal to the amount of money patients would have earned had they not been injured or ill.[36,86] Thus, productivity losses are generated during the complete duration of absence from paid work.[1,18,33,36,67] The Friction Cost Approach attempts to adjust for the fact that workers might be (partially) replaced in case of long-term sickness absence or premature mortality by truncating productivity losses at the friction period.[36,86] The friction period is the period needed to replace an absent sick worker and depends on the labour market, which means that its duration can differ between countries.[1,18,33,36,67] Both presenteeism and absenteeism can be valued using actual wage rates of patients, or age-, sex-, education-, and/or job-specific price weights.[36,98]

It is also possible that participants generate productivity losses related to unpaid work. Unpaid productivity losses are defined as losses due to an incapability to perform unpaid activities, such as volunteer work, household work, and education.[87] Unpaid productivity losses can be measured by asking patients to report the hours of unpaid work that they were unable to perform due to their health condition.[96] Unpaid productivity losses can be valued using the aforementioned proxy good costs and opportunity costs (see patient and family costs section). A more detailed explanation of the identification, measurement, and valuation of costs in the case study is presented in Box 3.

Box 3 | Identification, measurement, and valuation of costs in the case study

In the case study, the societal perspective was applied, and total costs included intervention costs, healthcare costs, patient and family costs, and lost productivity costs.[37] Intervention costs were estimated using a micro-costing approach. Information was gathered about the number of exercise sessions patients attended as well as the number of distributed information booklets, after which both items were valued using unit prices derived from the Brazilian physical therapy council (for the exercise sessions) and print shops (for printing of the booklets).[74] Healthcare costs included costs related to the use of medications and other health services. Information on the quantity of healthcare services consumed was collected during the trial using cost diaries developed by the researchers. Unit prices were derived from the SUS cost table.[77] Patient and family costs were collected by asking patients to report the number of kilometres travelled by car and/or the number of public transport tickets needed to get to the clinic as well as their expenses on over-the-counter and complementary medicines. Transportation by car was valued using Brazilian gasoline prices (R$0.23 per kilometre), and public transport was valued using the reference price of Sao Paulo city (£3.77 per trip). Informal care costs were not measured. Productivity losses included absenteeism from paid work and productivity losses related to unpaid work, while presenteeism was not included. Absenteeism from paid work was measured using a questionnaire and valued according to the Human Capital Approach using sex-specific price weights.[98] Productivity losses related to unpaid work were measured by asking patients the total number of hours of unpaid work that they were unable to perform due to their chronic low back pain. Unpaid productivity losses were valued using the same unit price as absence from paid work, because Brazilian reference prices for unpaid losses are lacking.

Adjusting costs for differential timing

In trial-based economic evaluations, it is common that unit prices are not available for the same year. Due to inflation, however, the price of goods and services will typically increase over time and consequently prices from different years are not directly comparable.[99] Therefore, all costs need to be converted to the same reference year using consumer price indices (CPI).[34,36,65,99] A more detailed explanation of converting prices to the same year using CPIs is presented in Box 4.

Box 4 | Converting prices using Consumer Price Indices (CPI)

In the case study, the reference year adopted was 2016.[37] All costs needed to be adjusted to the same reference year using the Brazilian consumer price indices below.[98]

Year	CPI
2014	3836.37
2015	4110.20
2016	4550.23

For this adjustment, we used the following formula: $Price_r = \left(\frac{Price_i}{CPI_i}\right) * \ CPI_r$
where $Price_i$ and CPI_i are the unit price and CPI of the index year and $Price_r$ and CPI_r are the unit price and CPI of the reference year. Thus, if we would like to convert the price of an exercise therapy session from Brazilian real in 2014 (R$70.50) to Brazilian real in 2016, we can do that as follows:[74]
Price2016 = (R$70.50/3836.37) * 4550.23 = R$83.61

Another phenomenon that should be considered in trial-based economic evaluations is that costs and effects are sometimes measured over more than one year. Since people have a preference to receive benefits today rather in the future, costs and effects occurring in the second and later years of follow-up need to be adjusted by converting them to their present value.[1,36,43,45,79,100,101,102] The appropriate discount rate differs between countries,[3,25,45,79,89] and may differ for costs and effects.[3,18,36,43,65,75,103] A more detailed explanation on how to apply discount rates is presented in Box 5.

Analysis and interpretation of an economic evaluation

Sample size

In trial-based economic evaluations, the sample size is usually estimated based on the anticipated clinically relevant difference in effect outcomes and not in costs. However, due to the right-skewed distribution of cost data (Figure 1), larger sample sizes are required to detect relevant differences in costs than in outcomes that follow a normal distribution. This right-skewed distribution of costs is caused by the fact that the majority of patients has relatively low costs, while few patients have high costs. The large sample sizes that would be required for cost differences are infeasible, and it may be considered unethical to continue recruiting patients into a trial beyond the point at which clinical superiority has been determined beyond reasonable doubt.[33,34,36,39,104,105,106] Consequently, trial-based economic evaluations are usually underpowered to detect relevant cost differences. To deal with this limitation, researchers are recommended to focus on estimation rather than hypothesis testing, that is on the relative magnitude of the cost and effect differences and their corresponding 95% confidence intervals (95% CIs), rather than on the corresponding p values.

Box 5 | Discounting of costs

In the case study, discounting of costs was not necessary due to the 12-month follow-up.37 Therefore, a hypothetical situation is used in the example below.

In a hypothetical study with a 3-year follow-up, a discount rate of 5% was applied to the cost of a manual therapy session.

Year	**Intervention costs**
1	R$5,00
2	R$10,00
3	R$15,00

The cost of a manual therapy session was R$70.00. Discounting of costs for the second and third year were conducted using the formula below:

$$P = \frac{F_0 + F_1}{(1+i)^1} + \frac{F_2}{(1+i)^2}$$

where P is the price of a manual therapy session in the present (i.e., present value), F_0 is the price of a manual therapy session in the first year, F_1 is the price of a manual therapy session in the second year, F_2 is the price of a manual therapy session in the third year, and is the discount rate (5%=0.05). Thus, if we would like to estimate the present value of the manual therapy session, we can do that as follows:

$$P = \frac{F_{2016} + F_{2017}}{(1+0.05)^1} + \frac{F_{2018}}{(1+0.05)^2}$$

$P = \frac{70+70}{(1+0.05)^1} + \frac{70}{(1+0.05)^2}$ = 70 + 66.67 + 63.49 = R$200.16

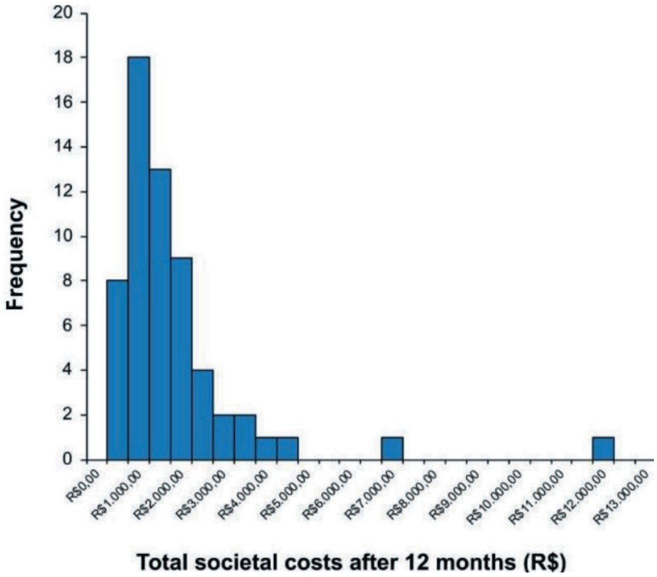


Figure 1 | Histogram showing the right-skewed distribution of societal costs in the case study. Most patients have relatively low costs, few patients have high costs, and costs cannot be lower than zero.

Statistical methods

Missing data

In trial-based economic evaluations, missing data may be a larger problem than in effectiveness evaluations because total costs are the sum of different cost components collected at various time points. If only one cost component is missing, total costs will be missing as well.[36,45] Missing data can be handled by simply deleting patients with missing values. This method, a so-called complete-case analysis, is not recommended, as it reduces a study's power and can lead to biased estimates. Moreover, not using all available data may even be considered unethical. Simple imputation methods, such as mean imputation and last observation carried forward, are also discouraged, because they do not account for the uncertainty related to filling in missing values.[32] Multiple imputation is currently considered a valid method for handling missing data in trial-based economic evaluations.[107,108,109,110,111] With multiple imputation, multiple datasets are created using multivariate techniques in which missing values are replaced by imputed values.[110,112] The imputed data sets are analysed separately to obtain a set of parameter estimates, which can then be pooled using Rubin's rules to obtain overall estimates, variances, and 95%CIs.[110,112]

Skewed costs

The skewed distribution of costs violates the assumption of standard statistical tests (e.g., linear regression and independent t test) that the data are normally distributed. A standard approach for analysing skewed data is to use standard non-parametric tests, such as a Mann-Whitney U test.[36] However, such non-parametric tests do not provide an estimate of the mean difference in costs between study groups, whereas decision-makers need this information to estimate the total budget needed to treat all patients with the new intervention.

Another commonly used approach is to transform skewed data, after which the data follow a normal distribution, such as a logarithmic transformation. However, statistical estimates based on log-transformation are difficult to interpret, because the mean differences between groups are expressed on a log-scale.[84] Back transformation will result in an estimate of the percentage of difference in costs between groups, instead of a mean difference.[84] Therefore, the ISPOR RCT-CEA guideline recommends the use of non-parametric bootstrapping to deal with the highly skewed nature of cost data.[32] With this approach, statistical analyses are based on repeated samples with replacement drawn from the original sample of the study (observed data).[36,113] In summary, a sample of patients that is equal in size to the study group is repeatedly randomly drawn with replacement from the intervention and control groups, separately.[36,113] Each resulting dataset is called a bootstrap sample and can be considered the mathematical equivalent of a replication of the study.[36,113] Each bootstrap sample will differ from the original sample, because the replacement of patients means that a specific observation can be included more than once in a bootstrap sample. Then, the statistic of interest is estimated (e.g., difference in costs) for every bootstrap sample.[36,113] Based on the central limit theorem, the distribution of the statistic of interest over the large number of bootstrap samples will approximate the normal distribution. The bootstrap samples can therefore be used to estimate confidence intervals (CIs).[36,113] Several methods are available to estimate CIs, including the percentile and bias-corrected and accelerated (BCA) bootstrap.[113,114] In the percentile method, 95%CIs are obtained by finding the values from the bootstrap distribution that correspond to the percentiles indicating the upper (97.5%) and lower (2.5%) bound of the CI.[113,114] In the BCA method, CIs are estimated using percentiles that are adjusted based on the skewness and bias of the data.[113,114] Of them, the BCA method is preferred. Research indicates that at least 2000 bootstrap samples are needed to produce reliable 95%CIs.[115]

Incremental cost-effectiveness ratio

The main outcome of interest in a trial-based economic evaluation is the incremental cost-effectiveness ratio (ICER).[1,18,33,34,36,39,116] The ICER is calculated by dividing the difference in mean costs between study groups (incremental costs = ΔCost) by the difference in mean effects (incremental effects = ΔEffect):[1,18,33,34,36,39,116]

$$ICER = \frac{Cost_{intervention} - Cost_{control}}{Effect_{intervention} - Effect_{control}} = \frac{\Delta Cost}{\Delta Effect}$$

ICERs can be interpreted as the amount of money that needs to be invested to gain one unit of effect extra. For example, in a CUA, the ICER reflects the incremental costs per QALY gained. ICERs on their own are generally hard to interpret. To illustrate, a negative ICER might represent two opposite situations: the intervention may be less expensive and more effective (a win-win situation, that is dominant) or more expensive and less effective (a lose-lose situation, that is dominated) than the comparator.[36] The cost-effectiveness plane (CE-plane) is often used to present ICERs. In the CE-plane, the difference in costs between groups is presented on the y-axis and the difference in effects on the x-axis, resulting in four quadrants.[18,33,34,36,117,118] An ICER located in the northeast quadrant indicates that the intervention is on average more effective and more costly than the comparator. An ICER located in the southeast quadrant indicates that the intervention is on average more effective and less costly (dominant) than the comparator. An ICER located in the southwest quadrant indicates that the intervention is on average less effective and less costly than the comparator. An ICER located in the northwest quadrant indicates that the intervention is on average less effective and more costly than the comparator (dominated).[18,33,34,36,117,118] An example of how to interpret ICERs is presented in Box 6.

CE-planes can also be used to provide an indication of the uncertainty surrounding the ICER point estimate.[18,33,34,36,117,118] Usually, the uncertainty surrounding a point estimate is given using 95%CIs. However, estimating 95%CIs around ICERs is not appropriate because the ICER is a ratio and therefore has an intractable distribution.[18,33,34,36,117,118] Therefore, non-parametric bootstrapping is typically used to estimate the uncertainty surrounding ICERs. Subsequently, all bootstrapped cost-effect pairs are plotted on the CE-plane. It is good practice to also show the percentage of bootstrapped cost-effect pairs per quadrant of the CE-plane as shown in Box 6.

The next step is to decide whether the intervention is cost-effective in comparison with control. When the ICER and most of the uncertainty is located in the southeast quadrant of the CE-plane the intervention can be considered dominant over control and, thus, cost-effective, while the northwest quadrant indicates the opposite. However, in the other two quadrants, i.e., the northeast quadrant and the southwest quadrant, the decision is less clear and depends on the amount of money decision-makers are willing to pay per unit of effect gained. That is, an ICER that is located in the northeast quadrant can only be considered cost-effective if the ICER is smaller than some predefined Willingness-to-pay (WTP) value, also known as WTP threshold.[18]

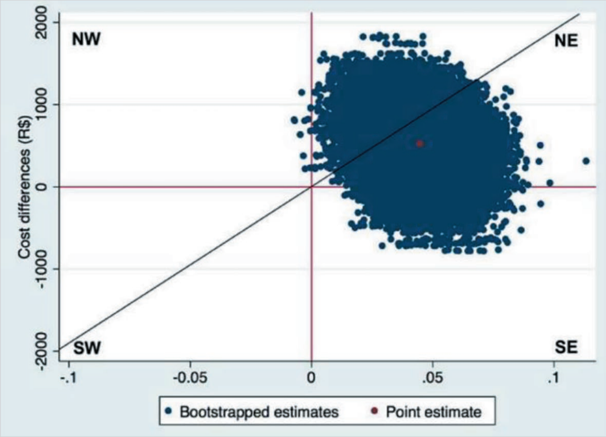
WTP thresholds are mainly defined for QALYs, while WTP thresholds for other important clinical outcomes in physical therapy research are lacking (i.e., pain intensity or disability).[119] In the United Kingdom, the WTP threshold is £20 000 to £30 000 per QALY gained, while in the Netherlands the WTP threshold ranges between €10 000 and €80 000 per QALY gained depending on the severity of the health condition.[119,120] In Brazil, there is no formal WTP threshold. The Brazilian guideline therefore recommends using the WTP threshold proposed by World Health Organization,[121] which is based on the per capita Gross Domestic Product (GDP) and varies from R$34 500 to R$103 600. However, the true value of the WTP, and the WTP threshold for disease-specific outcomes, are often not known. Therefore, the probability of an intervention being cost-effective at different WTP values is presented in a Cost-Effectiveness Acceptability Curve (CEAC).[18,33,36,122] The y-axis of a CEAC represents the probability of cost-effectiveness and the x-axis represents different WTP thresholds, that is the proportion of cost-effect pairs falling below a specific WTP threshold.[36,117,122,123] An example of a CEAC can be found in Box 6.

Sensitivity analyses

Trial-based economic evaluations are typically conducted in the context of incomplete information and uncertainty. Therefore, many assumptions need to be made. Sensitivity analyses should be performed to assess the robustness of the results to deviations from these assumptions.[36,65] Examples of sensitivity analyses are assessment of how study results would change when using a different perspective (e.g., healthcare perspective versus societal perspective), a different questionnaire for estimating QALYs (e.g., SF-6D versus EQ-5D), or a different strategy for handling missing data (e.g., complete-case analysis versus multiple imputation). In the case study, we performed two sensitivity analysis.[37] The first sensitivity analysis was performed from a healthcare perspective, and the second sensitivity analysis was performed per protocol, in which, only patients who attended more than 75% of the exercise sessions were included in the analyses.[37]

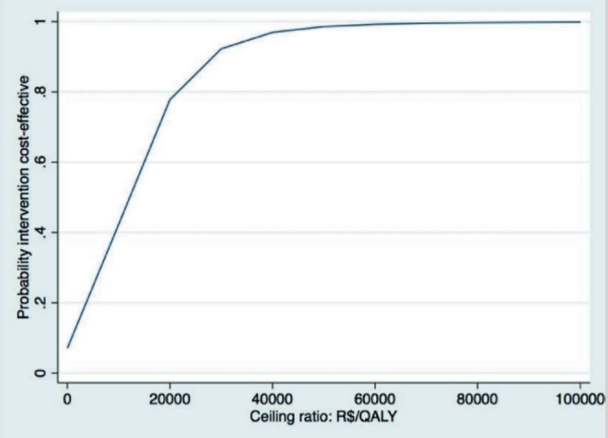
Box 6 | Interpretation of trial-based economic evaluation results

In the case study, analyses were performed according to the intention-to-treat principle and multiple imputation was used for handling missing data.[37] Non-parametric bootstrapping was used with 5000 replications and 95% CIs around cost and effect differences were estimated using the BCA approach.[35]



NW: northwest quadrant; SW: southwest quadrant; NE: northeast quadrant; SE: southeast quadrant

The Figure above shows the CE-plane for QALYs in the case study with a diagonal line representing a hypothetical WTP threshold.[37] The red dot (in the centre of the cloud) represents the point estimate of the ICER (ΔCost/ΔEffect = R$525/0.04 = 12 508 R$/QALY), and the blue dots represent the 5000 bootstrapped cost-effect pairs. Thus, on average, exercise therapy incurred an additional cost of R$12 506 per QALY gained compared to control. Furthermore, most of the bootstrapped cost-effect pairs are located in the northeast quadrant (92.5%), followed by the southeast quadrant (7.5%), northwest quadrant (<0.1%), and southwest quadrant (0.0%). This indicates that exercise therapy is most likely to be more costly and more effective than advice. The diagonal line in the CE-plane represents a WTP threshold of 22 727 R$/QALY gained. This line divides the cost-effectiveness plane into a cost-effective part (i.e. below the line) and a non-cost-effective part (i.e. above the line). Hence, ICERs located below this line can be considered cost-effective and ICER located above this line cannot be considered cost-effective.[34,36,117,118]



The Figure above shows the CEAC for QALYs gained of the case study.[37] CEAC shows the probabilities of cost-effectiveness on the y-axis and different WTP thresholds on the x-axis. We use the WTP threshold of R$45 455 per QALY gained (i.e., £20 000 per QALY gained) defined by the United Kingdom NICE here to evaluate whether exercise therapy was cost-effective compared to advice.119 This threshold was chosen, because a formal WTP threshold is not available for Brazil. At this WTP value, the probability of cost-effectiveness of exercise therapy compared to control was 95%. Based on these results, we concluded that exercise therapy is likely to be a cost-effective intervention compared to advice.

Discussion

Because musculoskeletal disorders are associated with a high burden to society and physical therapy interventions are important in the treatment of musculoskeletal disorders, information on the cost-effectiveness of such interventions has been increasingly requested by decision-makers. Collaborations between physical therapists, researchers, and health economists are needed to generate high quality evidence on the cost-effectiveness of physical therapy interventions. In this masterclass, we discussed the most important aspects that need to be considered when performing a trial-based economic evaluation, that is the perspective, the time horizon, the identification, measurement, and valuation of costs and effects, and methods used for costs and effect comparisons, missing data and uncertainty.

Recently, the WHO-EU "Research Agenda for Health Economic Evaluation" project identified three important challenges to economic evaluations in musculoskeletal health that, if addressed, could improve the use of health economic evidence in practice.[2,26] These challenges include the reporting quality of trial-based economic evaluations, their handling of uncertainty, and the issue of publication bias.[26] An increased use of reporting guidelines for trial-based economic evaluations, such as the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement may help improve the reporting quality of economic evaluations.[26] The CHEERS can be used by researchers as a guide when designing and reporting an economic evaluation and by Journal editors to assess the quality of such studies during the peer-review process. The handling of uncertainty may be improved by encouraging researchers to estimate the precision of the cost-effectiveness estimates using non-parametric bootstrapping and to graphically illustrate the level of uncertainty in CE-planes and CEACs.[26] Finally, publication bias may be reduced by encouraging researchers to publish all of the intended economic evaluations, instead of only those of clinical trials with positive effect outcomes.[26]

Clinical trials offer a unique opportunity to prospectively collect patient-level cost and effect data, and therefore to assess the cost-effectiveness of physical therapy interventions. Nonetheless, several recent randomized controlled trials in physical therapy did not include an economic evaluation.[124,125,126] As the additional cost to conduct an economic evaluation alongside a clinical trial is only marginal, we encourage researchers in physical therapy to make the conduct of economic evaluations alongside clinical trials common practice.[34] When an economic evaluation is not added onto an effectiveness evaluation, the opportunity is lost to collect and analyse cost and effect data simultaneously, which might in turn lead to the potential implementation of effective interventions that are not cost-effective.[34] Because clinical practice is unruly, de-implementation of adopted intervention is difficult. Additionally, even though some researchers are of the opinion that economic evaluations should only be conducted and published after clinical effectiveness is established, we recommend researchers in physical therapy to always assess and report on the cost-effectiveness of their intervention, irrespective of the effectiveness results. Absence of a statistically significant cost and/or effect difference does not necessarily mean that an intervention is not cost-effective and/or cost-beneficial. That is, economic evaluations are about the joint distribution of costs and effects and high probabilities of cost-effectiveness can be found even when there are no significant differences in costs or effects. Moreover, reductions in costs can occur in the absence of clinical effects and could thus be missed if an economic evaluation is not performed.[34,36]

Decision-makers in healthcare are encouraged to use evidence from trial-based economic evaluations when deciding whether or not to implement and/or reimburse new interventions. In countries, such as Australia, the United Kingdom, and the Netherlands, the uptake of economic evaluation results in the healthcare decision-making process has increased considerably during the last decade(s).[3,127,128] Although this process is most clearly applied for new pharmaceuticals, other interventions are also more and more subject to such rigorous evaluations. For example, in The Netherlands a randomized controlled trial was reimbursed pending the decision whether or not to include radiofrequency denervation for patients with chronic low back pain in the Dutch basic health insurance package. The study showed that radiofrequency denervation was not effective, nor cost-effective, when added to a standardized exercise program. As a result, radiofrequency denervation was no longer covered by public health insurance in The Netherlands.[128,129]

Trial-based economic evaluations are considered the "gold standard" for making inferences about the cost-effectiveness of physical therapy interventions.[32,34,45,46] However, the large sample size required by the skewed costs is often unfeasible for trial-based economic evaluations and follow-up in randomized controlled trials is typically not long enough to detect all relevant differences in costs between study groups. Furthermore, (trial-based) economic evaluations are typically conducted in research settings that do not resemble actual clinical practice. Finally, the use of different perspectives limits the generalizability and transferability of results to other settings and/or countries.

Conclusions

Economic evaluations are increasingly being used in healthcare decision-making. Therefore, it is of utmost importance that their design, conduct, and analysis are state-of-the-art and that their interpretation is adequate. This masterclass may help physical therapists, researchers, and decision-makers in the field of physical therapy to better understand trial-based economic evaluations with the ultimate goal of increasing translation of the results of such studies into clinical practice. Table 1 describes a summary of recommendations for trial-based economic evaluation of musculoskeletal physical therapy interventions.

Table 1 | Summary of recommendations for trial-based economic evaluation of musculoskeletal physical therapy interventions

Design of economic evaluation

Perspective

The recommended perspective differs across countries. Because the applied perspective determines which cost categories are assessed and included in an economic evaluation, it should always be stated explicitly.

Time horizon

The time horizon should be long enough to allow for the assessment of all relevant costs and effects flowing from the intervention under study.

Identification, measurement, and valuation of effects

Most economic evaluations in physical therapy research include both a CEA (to inform healthcare providers) and a CUA (to inform healthcare decision-makers).

Identification, measurement, and valuation of costs

The resource use items that need to be included highly depend on the applied perspective, the interventions being evaluated, and the patient population. Once all relevant resource categories are identified, researchers should determine how to "cost" them. For that, ideally unit prices reflecting "opportunity costs" (i.e. the value of a resource in its most highly valued alternative use) are used. Moreover, the quantities of resources consumed as well as their respective unit prices are ideally reported separately so that readers can recalculate costs for their own setting.

Adjusting costs for differential timing

All costs from different years need to be converted to the same reference year using consumer price indices. Furthermore, costs and effects measured over more than one year need to be adjusted using discount rate.

Analysis and interpretation of an economic evaluation

Sample size

The sample size is usually estimated based on the anticipated clinically relevant difference in effect outcomes and not in costs.

Statistical methods

Missing data

Multiple imputation is currently considered the most valid method for handling missing data in trial-based economic evaluations.

Skewed costs

The skewed distribution of costs violates the assumption of standard statistical tests that the data are normally distributed. Non-parametric bootstrapping is the preferred method to deal with the highly skewed nature of cost data.

Incremental cost-effectiveness ratio

Incremental cost-effectiveness ratios (ICERs) can be interpreted as the amount of money that needs to be invested to gain one unit of effect extra. The cost-effectiveness plane is often used to present ICERs and can also be used to provide an indication of the uncertainty surrounding the ICER point estimate. Cost-effectiveness acceptability curves are used to provide an indication of the probability of an intervention being cost-effective at different willingness to pay thresholds.

Sensitivity analysis

Sensitivity analyses should be performed to assess the robustness of the results of an economic evaluation.

References

1. Drummond MF, SM TG. 3rd ed. Oxford University Press; USA: 2005. Methods for the Economic Evaluation of Health Care Programmes.
2. Tordrup D, Bertollini R. Consolidated research agenda needed for health economic evaluation in Europe. BMJ. 2014;349:g5228.
3. National Institute for Health and Clinical Excellence . NICE; 2008. NICE Guide to the Methods of Technology Appraisal.
4. Moraz G, Garcez ADS, Assis EMD, Santos JPD, Barcellos NT, Kroeff LR. Estudos de custo-efetividade em saúde no Brasil: uma revisão sistemática. Cienc Saude Colet. 2015;20:3211–3229.
5. Drake TL, Devine A, Yeung S, Day NP, White LJ, Lubell Y. Dynamic transmission economic evaluation of infectious disease interventions in low- and middle-income countries: A systematic literature review. Health Econ. 2016;25(Suppl 1):124–139.
6. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1545–1602.
7. Jackson T, Thomas S, Stabile V, Shotwell M, Han X, McQueen K. A systematic review and meta-analysis of the global burden of chronic pain without clear etiology in low- and middle-income countries: Trends in heterogeneous data and a proposal for new assessment methods. Anesth Analg. 2016;123(3):739–748.
8. Lalonde L, Choiniere M, Martin E, Berbiche D, Perreault S, Lussier D. Costs of moderate to severe chronic pain in primary care patients - a study of the ACCORD Program. J Pain Res. 2014;7:389–403.
9. Breivik H, Eisenberg E, O'Brien T. Openminds. The individual and societal burden of chronic pain in Europe: The case for strategic prioritisation and action to improve knowledge and availability of appropriate care. BMC Public Health. 2013;13:1229.
10. Dieleman JL, Cao J, Chapin A. US health care spending by payer and health condition, 1996-2016. JAMA. 2020;323(9):863–884.
11. Bevan S. Economic impact of musculoskeletal disorders (MSDs) on work in Europe. Best Pract Res Cl Rh. 2015;29(3):356–373.
12. Dagenais S, Caro J, Haldeman S. A systematic review of low back pain cost of illness studies in the United States and internationally. Spine J. 2008;8(1):8–20.
13. Salary converter: purchasing power parities. http://salaryconverter.nigelb.me/. Updated February. Accessed 2017.
14. Carregaro RL, da Silva EN, van Tulder M. Direct healthcare costs of spinal disorders in Brazil. Int J Public Health. 2019;64(6):965–974.
15. Carregaro RL, Tottoli CR, Rodrigues DDS, Bosmans JE, da Silva EN, van Tulder M. Low back pain should be considered a health and research priority in Brazil: Lost productivity and healthcare costs between 2012 to 2016. PloS one. 2020;15(4)
16. Buchbinder R, Underwood M, Hartvigsen J, Maher CG. The Lancet Series call to action to reduce low value care for low back pain: an update. Pain. 2020;161(Suppl 1):S57–S64.
17. Foster NE, Anema JR, Cherkin D. Prevention and treatment of low back pain: Evidence, challenges, and promising directions. Lancet. 2018;391(10137):2368–2383.
18. van der Roer N, Boos N, van Tulder MW. Economic evaluations: A new avenue of outcome assessment in spinal disorders. Eur Spine J. 2006;15(Suppl 1):S109–S117.
19. Decimoni TC, Leandro R, Rozman LM. Systematic review of health economic evaluation studies developed in brazil from 1980 to 2013. Front Public Health. 2018;6:52.

20. Kanters TA, Bouwmans CAM, van der Linden N, Tan SS, Hakkaart-van Roijen L. Update of the Dutch manual for costing studies in health care. PloS one. 2017;12(11)

21. Rede Brasileira de Avaliação de Tecnologias em Saúde. http://rebrats.saude.gov.br. Published 2020. Accessed.

22. Rede Brasileira de Avaliação de Tecnologias em Saúde (REBRATS). Diretriz de Avaliação Econômica. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. Departamento de Gestão e Incorporação de Tecnologias em Saúde. REBRATS. http://rebrats.saude.gov.br/diretrizes-metodologicas. Updated 2013. Accessed 2017.

23. Peter WF, Jansen MJ, Hurkmans EJ. Physiotherapy in hip and knee osteoarthritis: development of a practice guideline concerning initial assessment, treatment and evaluation. Acta Reumatol Port. 2011;36(3):268–281.

24. Childs JD, Cleland JA, Elliott JM. Neck pain: clinical practice guidelines linked to the international classification of functioning, disability, and health from the orthopedic section of the american physical therapy association. J Orthop Sports Phys Ther. 2008;38(9):A1–A34.

25. Qaseem A, Wilt TJ, McLean RM, Forciea MA. Clinical Guidelines Committee of the American College of P. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. Ann Intern Med. 2017;166(7):514–530.

26. van Dongen JM, Ketheswaran J, Tordrup D, Ostelo R, Bertollini R, van Tulder MW. Health economic evidence gaps and methodological constraints in low back pain and neck pain: results of the research agenda for health economic evaluation (RAHEE) project. Best Pract Res Cl Rh. 2016;30(6):981–993.

27. Roseboom KJ, van Dongen JM, Tompa E, van Tulder MW, Bosmans JE. Economic evaluations of health technologies in Dutch healthcare decision-making: a qualitative study of the current and potential use, barriers, and facilitators. BMC Health Serv Res. 2017;17(1):89.

28. Eddama O, Coast J. A systematic review of the use of economic evaluation in local decision-making. Health Policy. 2008;86(2-3):129–141.

29. Hoffmann C, Graf von der Schulenburg JM. The influence of economic evaluation studies on decision making. A European survey. The EUROMET group. Health Policy. 2000;52(3):179–192.

30. Hoffmann C, Stoykova BA, Nixon J, Glanville JM, Misso K, Drummond MF. Do health-care decision makers find economic evaluations useful? The findings of focus group research in UK health authorities. Value Health. 2002;5(2):71–78.

31. Zwart-van Rijkom JE, Leufkens HG, Busschbach JJ, Broekmans AW, Rutten FF. Differences in attitudes, knowledge and use of economic evaluations in decision-making in The Netherlands. The Dutch results from the EUROMET Project. PharmacoEconomics. 2000;18(2):149–160.

32. Ramsey SD, Willke RJ, Glick H. Cost-effectiveness analysis alongside clinical trials II-An ISPOR Good Research Practices Task Force report. Value Health. 2015;18(2):161–172.

33. Korthals-de Bos I, van Tulder M, van Dieten H, Bouter L. Economic evaluations and randomized trials in spinal disorders: principles and methods. Spine. 2004;29(4):442–448.

34. Petrou S, Gray A. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. BMJ (Clinical research ed) 2011;342:d1548.

35. van Dongen JM, El Alili M, Varga AN. What do national pharmacoeconomic guidelines recommend regarding the statistical analysis of trial-based economic evaluations? Expert Rev Pharmacoecon Outcomes Res. 2019:1–11.

36. van Dongen JM, van Wier MF, Tompa E. Trial-based economic evaluations in occupational health: principles, methods, and recommendations. J Occup Environ Med. 2014;56(6):563–572.

37. Miyamoto GC, Franco KFM, van Dongen JM. Different doses of Pilates-based exercise therapy for chronic low back pain: a randomised controlled trial with economic evaluation. Br J Sports Med. 2018

38. Chiarotto A, Boers M, Deyo RA. Core outcome measurement instruments for clinical trials in nonspecific low back pain. Pain. 2018;159(3):481–495.

39. Gray A, Clarke PM, Wolstenholme JL, Wordsworth S. Oxford University Press; 2010. Applied Methods of Cost-Effectiveness Analysis in Healthcare.

40. Petrou S, Gray A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. BMJ. 2011;342:d1766.

41. Evers SM, Hiligsmann M, Adarkwah CC. Risk of bias in trial-based economic evaluations: identification of sources and bias-reducing strategies. Psychol Health. 2015;30(1):52–71.

42. Roberts M, Russell LB, Paltiel AD, Chambers M, McEwan P, Krahn M. Conceptualizing a model: a report of the ISPOR-SMDM modeling good research practices task force -2. Value Health. 2012;15(6):804–811.

43. Siebert U, Alagoz O, Bayoumi AM. State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task force-3. Value Health. 2012;15(6):812–820.

44. Eddy DM, Hollingworth W, Caro JJ. Model transparency and validation: a report of the ISPOR-SMDM modeling good research practices task force-7. Med Decis Making. 2012;32(5):733–743.

45. Ramsey S, Willke R, Briggs A. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA task force report. Value Health. 2005;8(5):521–533.

46. Howards PP. An overview of confounding. Part 1: the concept and how to address it. Acta Obstet et Gynecol Scand. 2018;97(4):394–399.

47. Brouwer WBF, van Exel NJA, Baltussen RMPM, Rutten FFH. A dollar is a dollar is a dollar—or is it? Value Health. 2006;9(5):341–347.

48. Schreijenberg M, Luijsterburg PA, Van Trier YD. Efficacy of paracetamol, diclofenac and advice for acute low back pain in general practice: design of a randomized controlled trial (PACE Plus) BMC Musculoskelet Disord. 2017;18(1):56.

49. Miyamoto GC, Lin CC, Cabral CMN, van Dongen JM, van Tulder MW. Cost-effectiveness of exercise therapy in the treatment of non-specific neck pain and low back pain: A systematic review with meta-analysis. Br J Sports Med. 2019;53(3):172–181.

50. Pinto D, Robertson MC, Hansen P, Abbott JH. Cost-effectiveness of nonpharmacologic, nonsurgical interventions for hip and/or knee osteoarthritis: systematic review. Value Health. 2012;15(1):1–12.

51. Tsertsvadze A, Clar C, Court R, Clarke A, Mistry H, Sutcliffe P. Cost-effectiveness of manual therapy for the management of musculoskeletal conditions: a systematic review and narrative synthesis of evidence from randomized controlled trials. J Manipulative Physiol Ther. 2014;37(6):343–362.

52. Nunes Cabral CM, Miyamoto GC, Moura Franco KF, Bosmans JE. Economic evaluations of educational, physical, and psychological treatments for fibromyalgia: A systematic review with meta-analysis. Pain. 2021

53. Chiarotto A, Ostelo RW, Turk DC, Buchbinder R, Boers M. Core outcome sets for research and clinical practice. Braz J Phys Ther. 2017;21(2):77–84.

54. Page MJ, McKenzie JE, Green SE. Core domain and outcome measurement sets for shoulder pain trials are needed: systematic review of physical therapy trials. J Clin Epidemiol. 2015;68(11):1270–1281.

55. Choy EH, Arnold LM, Clauw DJ. Content and criterion validity of the preliminary core dataset for clinical trials in fibromyalgia syndrome. J Rheumatol. 2009;36(10):2330–2334.

56. Smith TO, Mansfield M, Hawker GA. Uptake of the OMERACT-OARSI hip and knee osteoarthritis core outcome set: review of randomized controlled trials from 1997 to 2017. J Rheumatol. 2019;46(8):976–980.

57. Murray CJ, Acharya AK. Understanding DALYs (disability-adjusted life years) J Health Econ. 1997;16(6):703–730.

58. Sassi F. Calculating QALYs, comparing QALY and DALY calculations. Health Policy Plan. 2006;21(5):402–408.

59. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions. Health Utilities Index. PharmacoEconomics. 1995;7(6):503–520.

60. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35(11):1095–1108.

61. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21(2):271–292.

62. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. Br Med Bull. 2010;96:5–21.

63. Cruz LN, Camey SA, Hoffmann JF. Estimating the SF-6D value set for a population-based sample of Brazilians. Value Health. 2011;14(5 Suppl 1):S108–S114.

64. Briggs AH, O'Brien BJ. The death of cost-minimization analysis? Health Econ. 2001;10(2):179–184.

65. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ economic evaluation working party. BMJ. 1996;313(7052):275–283.

66. Drummond M, Manca A, Sculpher M. Increasing the generalizability of economic evaluations: Recommendations for the design, analysis, and reporting of studies. Int J Technol Assess Health Care. 2005;21(2):165–171.

67. Uegaki K, de Bruijne MC, Anema JR, van der Beek AJ, van Tulder MW, van Mechelen W. Consensus-based findings and recommendations for estimating the costs of health-related productivity loss from a company's perspective. Scand J Work Environ Health. 2007;33(2):122–130.

68. Beemster TT, van Velzen JM, van Bennekom CAM, Reneman MF, Frings-Dresen MHW. Test-retest reliability, agreement and responsiveness of Productivity Loss (iPCQ-VR) and Healthcare Utilization (TiCP-VR) Questionnaires for sick workers with chronic musculoskeletal pain. J Occup Rehabil. 2019;29(1):91–103.

69. Hoefman RJ, Van Exel NJA, Brouwer WBF. iVICQ. iMTA Valuation of Informal Care Questionnaire, version 1.1. http://www.bmg.eur.nl/english/imta/publications/questionnaires_manuals/ivicq/. Published 2013. Accessed.

70. Database of Instruments for Resource Use Measurement. http://www.dirum.org/. Published 2020. Accessed.

71. Goossens ME, Rutten-van Molken MP, Vlaeyen JW, van der Linden SM. The cost diary: a method to measure direct and indirect costs in cost-effectiveness research. J Clin Epidemiol. 2000;53(7):688–695.

72. van den Brink M, van den Hout WB, Stiggelbout AM, Putter H, van de Velde CJ, Kievit J. Self-reports of health-care utilization: diary or questionnaire? Int J Technol Assess Health Care. 2005;21(3):298–304.

73. Tompa E, Culyer AJ, Dolinschi J. Oxford University Press; New York: 2008. Economic Evaluation of Interventions for Occupational Health and Safety: Developing Good Practive.

74. COFFITO: Conselho Federal de Fisioterapia e Terapia Ocupacional. http://coffito.gov.br/nsite/wp-content/uploads/2016/08/CartilhadeValoriza%C3%A7%C3%A3oProfissional_Fisioterapia_vers%C3%A3o01_04_2016.pdf. Updated August. Accessed 2017.

75. Hakkaart-van Roijen L, Tan S, Bouwmans C. College voor zorgverzekeringen; Diemen: 2010. Handleiding voor Kostenonderzoek. Methoden en Standaardkostprijzen voor Economische Evaluaties in de Gezondheidszorg.

76. Oostenbrink JB, Koopmanschap MA, Rutten FF. Standardisation of costs: the Dutch Manual for Costing in economic evaluations. PharmacoEconomics. 2002;20(7):443–454.

77. SIGTAP - Sistema de Gerenciamento da Tabela de Procedimentos. DATASUS. sigtap.datasus.gov.br. Updated April. Accessed 2017.

78. Dakin H, Abangma G, Wordsworth S. What is the value of collecting detailed costing data in clinical trials? Trials. 2011;12(1):A42.

79. Rede Brasileira de Avaliação de Tecnologias em Saúde (REBRATS). Diretriz metodológica: Estudos de microcusteio aplicados a avaliações econômicas em saúde. REBRATS.https://rebrats.saude.gov.br/images/Documentos/Diretriz_Metodologica_Estudos_de_Microcusteio_Aplicados_a_Avaliacoes_Economicas_em_Saude.pdf. Accessed 2020.

80. Frick KD. Microcosting quantity data collection methods. Med Care. 2009;47(7 Suppl 1):S76–S81.

81. Chapel JM, Wang G. Understanding cost data collection tools to improve economic evaluations of health interventions. Stroke Vasc Neurol. 2019;4(4):214–222.

82. Clement Nee Shrive FM, Ghali WA, Donaldson C, Manns BJ. The impact of using different costing methods on the results of an economic evaluation of cardiac care: Microcosting vs gross-costing approaches. Health Econ. 2009;18(4):377–388.

83. Raftery J. Costing in economic evaluation. BMJ. 2000;320(7249) 1597-1597.

84. Glick H.A., Doshi J.A., Sonnad S.S., D. P. Oxford University Press; New York: 2007. Economic Evaluations in Clinical Trials.

85. van den Berg B, Brouwer WB, Koopmanschap MA. Economic valuation of informal care. An overview of methods and applications. Eur J Health Econ. 2004;5(1):36–45.

86. Zhang W, Bansback N, Anis AH. Measuring and valuing productivity loss due to poor health: a critical review. Soc Sci Med. 2011;72(2):185–192.

87. Krol M, Brouwer W, Rutten F. Productivity costs in economic evaluations: past, present, future. PharmacoEconomics. 2013;31(7):537–549.

88. Brouwer WB, Koopmanschap MA, Rutten FF. Productivity costs in cost-effectiveness analysis: numerator or denominator: a further discussion. Health Econ. 1997;6(5):511–514.

89. Kessler RC, Barber C, Beck A. The world health organization health and work performance questionnaire (HPQ) J Occup Environ Med. 2003;45(2):156–174.

90. Brouwer WB, Koopmanschap MA, Rutten FF. Productivity losses without absence: measurement validation and empirical evidence. Health Policy. 1999;48(1):13–27.

91. Koopmanschap MA. PRODISQ: a modular questionnaire on productivity and disease for economic evaluation studies. Expert Rev Pharmacoecon Outcomes Res. 2005;5(1):23–28.

92. Meerding WJ, IJ W, Koopmanschap MA, Severens JL, Burdorf A. Health problems lead to considerable productivity loss at work among workers with high physical load jobs. J Clin Epidemiol. 2005;58(5):517–523.

93. Lerner D, Amick BC, 3rd, Lee JC. Relationship of employee-reported work limitations to work productivity. Med Care. 2003;41(5):649–659.

94. Lerner D, Amick BC, 3rd, Rogers WH, Malspeis S, Bungay K, Cynn D. The Work Limitations Questionnaire. Med Care. 2001;39(1):72–85.

95. Lerner D, Reed JI, Massarotti E, Wester LM, Burke TA. The Work Limitations Questionnaire's validity and reliability among patients with osteoarthritis. J Clin Epidemiol. 2002;55(2):197–208.

96. Bouwmans C, Krol M, Severens H, Koopmanschap M, Brouwer W, Hakkaart-van Roijen L. The iMTA productivity cost questionnaire: a standardized instrument for measuring and valuing health-related productivity losses. Value Health. 2015;18(6):753–758.

97. Kigozi J, Jowett S, Lewis M, Barton P, Coast J. The Estimation and inclusion of presenteeism costs in applied economic evaluation: a systematic review. Value Health. 2017;20(3):496–506.

98. Instituto Brasileiro de Geografia e Estatística - IBGE. www.ibge.gov.br. Updated April. Accessed 2017.

99. Turner HC, Lauer JA, Tran BX, Teerawattananon Y, Jit M. Adjusting for inflation and currency changes within health economic studies. Value Health. 2019;22(9):1026–1032.

100. Attema AE, Brouwer WBF, Claxton K. Discounting in economic evaluations. PharmacoEconomics. 2018;36(7):745–758.

101. Gravelle H, Smith D. Discounting for health effects in cost-benefit and cost-effectiveness analysis. Health Econ. 2001;10(7):587–599.

102. Goossens ME, Evers SM, Vlaeyen JW, Rutten-van Molken MP, van der Linden SM. Principles of economic evaluation for interventions of chronic musculoskeletal pain. Eur J Pain. 1999;3(4):343–353.

103. Brouwer WB, Niessen LW, Postma MJ, Rutten FF. Need for differential discounting of costs and health effects in cost effectiveness analyses. BMJ. 2005;331(7514):446–448.

104. Barber JA, Thompson SG. Analysis and interpretation of cost data in randomised controlled trials: review of published studies. BMJ. 1998;317(7167):1195–1200.

105. Thompson SG, Barber JA. How should cost data in pragmatic randomised trials be analysed? BMJ. 2000;320(7243):1197–1200.

106. Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. Am J Epidemiol. 2005;161(2):105–110.

107. Gabrio A, Mason AJ, Baio G. Handling missing data in within-trial cost-effectiveness analysis: a review with future recommendations. PharmacoEconomics. 2017;1(2):79–97.

108. Noble SM, Hollingworth W, Tilling K. Missing data in trial-based cost-effectiveness analysis: The current state of play. Health Econ. 2012;21(2):187–200.

109. Diaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. Clin Trials. 2014;11(5):590–600.

110. Sterne JA, White IR, Carlin JB. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

111. Leurent B, Gomes M, Faria R, Morris S, Grieve R, Carpenter JR. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: A tutorial. PharmacoEconomics. 2018;36(8):889–901.

112. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377–399.

113. Barber JA, Thompson SG. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. Stat Med. 2000;19(23):3219–3236.

114. Kelley K. The effects of nonnormal distributions on confidence intervals around the standardized mean difference: bootstrap and parametric confidence intervals. Educ Psychol Meas. 2005;65(1):51–69.

115. Chaudhary MA, Stearns SC. Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. Stat Med. 1996;15(13):1447–1458.

116. Robinson R. Economic evaluation and health care. What does it mean? BMJ. 1993;307(6905):670–673.

117. Briggs AH, O'Brien BJ, Blackhouse G. Thinking outside the box: recent advances in the analysis and presentation of uncertainty in cost-effectiveness studies. Ann Rev Public Health. 2002;23:377–401.

118. Black WC. The CE plane: a graphic representation of cost-effectiveness. Med Decis Making. 1990;10(3):212–214.

119. Appleby J, Devlin N, Parkin D. NICE's cost effectiveness threshold. BMJ. 2007;335(7616):358–359.

120. Zwaap J, Knies S, van der Meijden C, Staal P, van der Heiden L. Ministerie van VWS; In Z. Nederland (Ed.) ed.Diemen: 2015. Kosteneffectiviteit in de Praktijk.

121. World Health Organization, Marseille E, Larson B, Kazi D, Kahn J, Rosen S. Thresholds for the cost–effectiveness of interventions: Alternative approaches. https://www.who.int/bulletin/volumes/93/2/14-138206/en/. Published 2019. Accessed.

122. Fenwick E, Marshall DA, Levy AR, Nichol G. Using and interpreting cost-effectiveness acceptability curves: an example using data from a trial of management strategies for atrial fibrillation. BMC Health Serv Res. 2006;6:52.

123. Fenwick E, O'Brien BJ, Briggs A. Cost-effectiveness acceptability curves–facts, fallacies and frequently asked questions. Health Econ. 2004;13(5):405–415.

124. Kashfi P, Karimi N, Peolsson A, Rahnama L. The effects of deep neck muscle-specific training versus general exercises on deep neck muscle thickness, pain and disability in patients with chronic non-specific neck pain: protocol for a randomized clinical trial (RCT) BMC Musculoskelet Disord. 2019;20(1):540.

125. Silva NC, Silva MC, Guimaraes MG, Nascimento MBO, Felicio LR. Effects of neuromuscular training and strengthening of trunk and lower limbs muscles in women with Patellofemoral Pain: a protocol of randomized controlled clinical trial, blinded. Trials. 2019;20(1):586.

126. Pustivsek S, Sarabon N. Integral movement therapy versus local movement therapy approach in patients with idiopathic chronic low-back pain: study protocol for a randomized controlled trial. Trials. 2019;20(1):69.

127. Australian Government, Department of Health. Health Technology Assessment (HTA). http://www.health.gov.au/internet/hta/publishing.nsf/Content/commonwealth-1. Published 2020. Accessed 24 Jul 2020.

128. Ministerie van Volksgezondheid, Welzijn en Sport. ZIN. Zorginstituut Nederland - Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg.https://www.zorginstituutnederland.nl/publicaties/publicatie/2016/02/29/richtlijn-voor-het-uitvoeren-van-economische-evaluaties-in-de-gezondheidszorg. Published 2016. Accessed 28 Mar 2018.

129. Juch JNS, Maas ET, Ostelo R. Effect of radiofrequency denervation on pain intensity among patients with chronic low back pain: the Mint randomized clinical trials. JAMA. 2017;318(1):68–81.

CHAPTER 10

General discussion

Summary

Allocating scarce healthcare resources as efficiently as possible has become a high priority for healthcare systems worldwide.[1–6] Economic evaluations can provide information about the “value for money” of health technologies and aid the healthcare decision-making process.[7] However, to ensure that healthcare resources can indeed be distributed efficiently, it is of utmost importance for economic evaluations to be “*scrupulous*”. Scrupulousness means that research is conducted “*using methods that are scientific or scholarly and exercising the best possible care in designing, undertaking, reporting and disseminating research*”.[8] In the area of trial-based economic evaluations, various gaps in knowledge exist that make it unclear how certain design and analysis steps can be scrupulously conducted. Three of these gaps in knowledge were addressed in this thesis, namely:

- The impact of using crosswalks on healthcare decision-making (*Chapters 2, 3*, and *4*);
- The prediction of health-related quality of life from a condition-specific patient-reported outcome measure (PROM) (*Chapters 5* and *6*);
- The handling of missing data in trial-based economic evaluations (*Chapter 7*).

Addressing these knowledge gaps alone will not improve the conduct of trial-based economic evaluations. Therefore, this thesis also included two tutorial papers on how to design, conduct, analyse, and interpret trial-based economic evaluations (*Chapters 8* and *9*).

Below, the main findings of this thesis will be discussed and compared with existing literature, after which some methodological considerations are discussed and implications for practice and future research are summarized.

Main Findings and comparison with the literature

The impact of using crosswalks on healthcare decision-making

In *Chapter 2*, we compared the EQ-5D-5L crosswalk[9] with the 5L value sets for England,[10] the Netherlands,[11] and Spain[12] and explored how cost-utility outcomes differed between these utility scoring methods in two case studies on depression and diabetes. Our findings showed that the 5L value set produced utility values that were on average higher than the EQ-5D-5L crosswalk set in both case studies for England and Spain, whereas the opposite was true for the Netherlands. Similar to our findings for England, Mulhern et al.[13] and Camacho et al.,[14] both found that the England 5L value set produced utility values that are on average 0.08 points higher than the U.K. EQ-5D-5L crosswalk across 12 health conditions, including depression and diabetes. In our two case studies, the observed differences in utility values between EQ-5D-5L crosswalks and 5L value sets did not translate into relevant differences in incremental QALYs. This is likely explained by the fact that differences between the utility scoring methods were similar in the intervention and control groups, thereby not affecting incremental outcomes, such as differences in QALYs. This observation is partly in line with the findings of Yang et al.[15] who found comparable incremental QALYs when comparing the crosswalk and value set for the Netherlands, but not for England. This discrepancy might be due to differences in the underlying study populations, as Yang et al. included severely ill patients, whereas moderately ill patients were included in our case studies. As a consequence of the lack of relevant differences in incremental QALYs that we observed, the interventions' probabilities of cost-effectiveness were not meaningfully affected by the use of the crosswalk or value set. To the best of our knowledge, previous studies assessing the impact of using different utility scoring methods on the probability of cost-effectiveness are lacking. Additional research is, therefore, needed to assess whether the current findings also apply to situations other than those explored in the case studies. That is, in populations with more severe health conditions and/or when interventions are more effective than the control.

Given the above, we decided to investigate whether the results of *Chapter 2* are generalizable to populations and/or interventions other than the ones included in the two case studies. This was done in *Chapter 3* by simulating a wide variety of trial-based economic evaluations based on published empirical data. These scenarios included a broader range of health conditions (i.e., depression, low back pain, osteoarthritis, and cancer) with different severity levels (i.e., mild, moderate, and severe), and interventions with different treatment effect sizes (i.e., small, medium, and large). We also investigated the impact of using the recently published reverse crosswalk (mapping from 3L to 5L)[16] compared to the 3L value set. Our findings indicated that the use of crosswalks, either 5L to 3L or 3L to 5L, instead of 5L or 3L value sets may impact cost-utility outcomes to such an extent that this may influence reimbursement decisions, particularly in scenarios with small treatment effect sizes. The differences in findings and conclusions between *Chapters 2* and *3* – among others – may be explained by differences in the strength and direction of the interventions' impact on costs and effects. That is, in *Chapter 2,* the case studies' interventions were on average "less effective" and "more costly" than control. In *Chapter 3*, however, we simulated scenarios with interventions that were "more effective" and "more costly" than control, which is a more likely scenario to occur in real-life reimbursement decisions. The findings of *Chapter* 3 also showed that the use of crosswalks is less likely to impact reimbursement decisions in countries that were used in the development of the crosswalk (i.e., Denmark, England, Italy, the Netherlands, Poland, and Scotland)[9]. This might suggest that crosswalks may not truly represent the preferences of specific populations other than those used for their development, especially when they have considerably different views on health-related quality of life. Further research into this issue is warranted.

In *Chapter 4*, we investigated whether the use of different country-specific value sets impacts ICERs and interventions' probabilities of cost-effectiveness. This was done using two case studies (i.e., one on low back pain and one on depression), both of which administered the EQ-5D-3L. For estimating utility values and QALYs, we used EQ-5D-3L value sets from 16 different countries and we also estimated 5L utility values by applying the crosswalk mapping approach.[9] In the case studies, we found that ICERs and probabilities of cost-effectiveness vary considerably across countries, which might in turn negatively impact the transferability of economic evaluation results from one country to another. These results also suggest that health state preferences are considerably affected by sociocultural differences between countries, which is in line with previous studies and underscores the importance of using country-specific value sets.[17–22] What differentiates *Chapter 4* from previous studies is that it also investigated the impact of country-specific EQ-5D value sets on the probability of an intervention being cost-effective compared to control. This is important because different utility values do not necessarily result in different trial-based economic evaluation results (e.g., ICER, probabilities of cost-effectiveness) and/or conclusions, as the impact of using a certain utility scoring method over another might be similar in the intervention and control group, thereby not affecting cost-utility outcomes. This was also seen in *Chapter 2*.

The prediction of health-related quality of life from a condition-specific patient-reported outcome measure

In *Chapter 5,* we assessed whether regression models could be used to predict EQ-5D-3L utility values from the Oswestry Disability Index (ODI)[23] for use in cost-effectiveness analyses among low back pain patients. For this purpose, we developed and validated six models using Ordinary Least Squares (OLS) and Tobit model: 1) OLS, with the total ODI score, 2) OLS, with the ODI item scores as continuous variables, 3) OLS, with the ODI item scores as ordinal variables, 4) Tobit model, with the total ODI score, 5) Tobit model, with the ODI item scores as continuous variables, 6) Tobit model, with the ODI item scores as ordinal variables. Two of the developed models (i.e., OLS and Tobit models with continuous ODI item scores) showed similar probabilities of cost-effectiveness compared to the Dutch 3L value set and were, therefore, considered adequate for the use in cost-effectiveness analyses. Based on these results we also concluded that the ODI can be used to predict LBP patients' EQ-5D-3L utility values when the aim is to perform a cost-utility analysis. However, as the two best-performing models had a relatively low absolute/relative fit and poor agreement between mapped and observed utility values, we do not recommend to use them for estimating HR-QoL for individual patients. The latter finding is in line with those of Carreon et al. [24] who concluded that individual patients' EQ-5D-3L utility values could not validly be predicted from their ODI scores.

Theoretical literature suggests that response mapping approaches perform better than regression models when predicting utility values from a condition-specific patient-reported outcome measure because they align the scales between instruments in such a way that the distributions of their responses are matched.[25–29] In *Chapter 6*, we, therefore, investigated whether response mapping results in better models, when trying to estimate EQ-5D-3L utility values from the ODI for use in cost-effectiveness analyses. Three response mapping approaches were employed: 1) a non-parametric approach, a 2) non-parametric approach excluding logical inconsistencies, and 3) an ordinal logistic regression. Results showed that the non-parametric approaches performed best as shown by a relatively low fit and wide – and clinically relevant – limits of agreement between observed and mapped utility values. However, these approaches did not perform better in terms of predicting individual patients' utility values than the best performing models from *Chapter 5*. Thus, based on our results response mapping approaches are not necessarily preferred over regression models for mapping PROMs to EQ-5D-3L. This finding is in contrast with the theoretical literature[25–29] that suggests that response mapping generally performs better than regression models to predict utility values. A possible explanation for this discrepancy might be that there were not many extreme scores (i.e., a ceiling effect) making regression to the mean less likely to occur, and that a relatively small proportion of the sample had a utility value of 1. Despite the low performance of the response mapping approaches in predicting HRQoL for individual patients, the differences in the probability of cost-effectiveness between observed and mapped values(1% to 4%) were relatively small. These differences were similar to those observed in *Chapter* 5 (1 to 5%).

The handling of missing data in trial-based economic evaluations

An important methodological challenge when performing a trial-based economic evaluation is that of missing data. Missing data is a problem because deleting cases with missing values from the analysis reduces a study's power and potentially biases cost-effectiveness estimates[30]. Advanced methods, such as Multiple Imputation (MI) and Longitudinal Linear Mixed-models (LLM), have been recommended to handle missing cost and effect data.[30–32] Nevertheless, the use of LLM and the added value of MI when using LLM to handle missing cost-effectiveness data has not been empirically tested. Therefore, in *Chapter 7*, we assessed whether MI is required prior to LLM when analysing longitudinal cost and effect data. For this purpose, 2000 datasets with baseline and follow-up cost and effect data were simulated with different proportions of missing data in follow-up costs and effects, and assuming a Missing At Random (MAR) mechanism. MAR occurs when the missing data is associated with observed variables, but not with unobserved variables.[33] Our findings suggest that LLM alone is appropriate for handling missing QALY data (i.e., it had an acceptable level of statistical performance), whereas the addition of MI prior to LLM improved the statistical performance considerably when handling missing cost data. At high levels of missing data (i.e., 50%) all methods had a relatively low level of statistical performance. Our results show that LLM alone is appropriate for handling missing QALY data under a MAR mechanism was in line with previous studies suggesting that when using LLM, MI of missing values is not necessary to obtain unbiased clinical effect estimates. Previous studies showed that the LLM performance for effects was good regardless of the missing data mechanism.[34,35] For costs, however, we found MI prior to

LLM to perform better than LLM alone under a MAR mechanism, which has not yet been shown before. The difference in performance between QALYs and costs is likely due to the fact that costs typically have higher levels of skewness, kurtosis, within-subject variability, and between-subject variability compared with QALYs. This is important because LMM assumes multivariate normality when analysing data, which is typically far from the case for cost data. Violations of the multivariate normality assumption may lead to non-convergence of the LLM model.[36] The MI model, on the other hand, uses the posterior distribution of data to predict missing values and, hence, does not require any distributional assumptions for the model to converge which may explain the superior performance of MI-LLM over LLM.[37,38]

Tutorials

In *Chapter 8*, we developed a tutorial to provide step-by-step guidance on how to combine statistical methods available in the literature to handle various methodological challenges inherent to trial-based economic evaluations using a ready-to-use R script. The methodological challenges addressed in this tutorial were missing data, correlated costs and effects, baseline imbalances, and the skewness of costs and effects. We explained the theoretical background of the described methods and illustrated how to apply them using a simulated trial-based economic evaluation. We additionally presented possible ways to extend the provided annotated R code and discussed the limitations of the approach chosen in this tutorial.

In *Chapter 9*, we developed a masterclass to discuss the best practices for the design, analysis, and interpretation of trial-based economic evaluations in the field of physical therapy. The masterclass was a project that was conducted in close collaboration with fellow colleagues from Brazil who identified the need to support Brazilian healthcare decision-makers with the interpretation of economic evaluations in the field of musculoskeletal disorders and helping them with translating the results of such studies into clinical practice.

Methodological considerations

Many of the methodological strengths and limitations of the studies included in this thesis have been discussed in *Chapters 2* to *7*. In addition, recommendations for good practice in trial-based economic evaluations have been provided in *Chapters 8* and *9*. Nonetheless, a selection of methodological challenges warrants further exploration and will be discussed below.

Simulated data

To assess the application, impact, and performance of statistical methods for trial-based economic evaluations we used both empirical and simulated data. Simulated datasets have the advantage over empirical datasets in providing us with knowledge about the "true value(s)" of the estimated quantities, which in turn allows evaluating how different methodological strategies compare to each other and the true value(s).[39] In other words, the performance (e.g., a measurement of how close predictions are to the true value) of competing methods can be assessed and compared.[39,40] This is what we did in *Chapter 7* to explore whether MI is required prior to LLM when analysing longitudinal cost and effect data. In our simulation study, we followed the latest recommendations of Morris et al.[39] to ensure that the design, conduct, and analysis of our simulation study were as optimal as possible. Amongst others, this meant that we calculated the number of simulated datasets needed to draw valid conclusions and that the simulated datasets resembled empirical data as closely as possible.[39]

In *Chapter 3*, we simulated data to assess the generalizability of results observed in a restrictive number of case studies to a wide range of other settings. We did so to assess whether our finding in *Chapter 2* that crosswalks and EQ-5D-5L value sets can be used interchangeably in the economic evaluations of two empirical case studies was generalizable to a broad range of scenarios, including different patient populations from different countries, and interventions with different treatment effect sizes. This turned out to be important because we found quite different trial-based economic evaluations results between crosswalks and EQ-5D value sets in a wide range of simulated scenarios compared to the relatively similar results that we found in the two empirical studies included in *Chapter 2*. However, a drawback of simulated data is that such data is always a simplification of reality, meaning that it may not represent the full complexity of real-world data. A strength is that simulations can help to scale down real-world problems. In doing so, simulations may help to assess the generalizability of results observed in a restrictive number of datasets and to get a better understanding of the problems and give directions on how to tackle them.[39,41]

Prediction modelling versus mapping approaches to estimate EQ-5D utility values

The fact that EQ-5D data might not always be available to estimate QALYs in economic evaluations of healthcare interventions motivates the use of regression-based prediction models and/or response mapping to estimate EQ-5D utility values from other measures of health outcomes.[42] With regression-based prediction models, utility values are directly predicted from a PROM score or PROM item responses.[42] Response mapping, on the other hand, is a specific type of mapping in which item responses between a source instrument (e.g., ODI, EQ-5D-5L) and a target instrument (e.g., EQ-5D-3L) are individually linked using an algorithm (e.g., a crosstabulation of responses). Based on the linked responses, the target instrument scores (e.g., utility values) can then be estimated.[27,42–45] Literature suggests that response mapping approaches might be better at preventing regression to the mean. Moreover, response mapping may also deal better with the well-known EQ-5D-3L ceiling effects compared to regression-based prediction models, because they align the scales between instruments so that the distributions of their responses are matched.[25–29] In *Chapters 5* and *6,* we assessed whether differences exist between mapped and observed utility values and found that mapping approaches can impact cost-utility outcomes to such an extent that this may influence reimbursement decisions. This indicates that the best practice is always to include a generic-based preference measure, like the EQ-5D, during the design phase of a study, and not to rely on regression-based prediction models and/or mapping approaches instead. However, when EQ-5D data is missing, *Chapters 5* and *6* showed that regression-based prediction models are preferred over mapping approaches for estimating utility values from the ODI for use in trial-based economic evaluations. For estimating utility values of individual low back pain patients, both approaches were not recommended due to their relatively low absolute and relative fit, that is, large root-mean-squared error (RMSE) and low explained variances (i.e., R squared values). However, low fit does not necessarily mean that the models cannot be used in the context of a cost-utility analysis, because bias is likely to be similar in the intervention and control groups, thereby not affecting incremental QALYs and CEACs. Further research into this area is warranted in other countries/populations, especially in patient populations with better health states and using the more sensitive version of the EQ-5D, i.e., the EQ-5D-5L.

Impact on the Cost-Effectiveness Acceptability Curve (CEAC)

In Chapters *2 to 7*, we not only assessed the impact of the different methodological strategies on cost and effect estimates of health technologies separately, but we also performed full trial-based economic evaluations to assess the impact of using different methodological strategies on cost-effectiveness outcomes, such as ICERs and cost-effectiveness acceptability curves (CEACs). We deemed this to be important, because performance measures, such as bias and RMSEs, only indicate the impact of using different methods on value sensitivity (e.g., their impact on the estimated cost and effect differences), but not on decision sensitivity (i.e., their impact on the eventual decision-making process). This is relevant because incremental cost and effect values are point estimates, which do not incorporate sampling uncertainty, whereas decision-makers need to know how certain they can be about the correctness of their decision. This decision uncertainty is commonly presented using cost-effectiveness acceptability curves (CEACs). Some authors argue that CEACs should primarily be used to decide whether more research on new technology is necessary given the to decide whether an intervention should be reimbursed or not, thus not in the decision-making process per se.[46,47] A recent study showed that sampling uncertainty as represented in CEACs is used to a limited extent only in reimbursement decisions in the Netherlands (not published yet). This may imply that scarce resources are wasted on technologies that are in reality not efficient. We, therefore, recommend decision-makers to not only use point estimates of ICERs, but to base their reimbursement decisions on CEACs.

The performance measures used in our study, bias, RMSE, and coverage probability, may not be sufficient to be used in the context of CEACs. Ideally, they should be supplemented by performance measures that are specifically developed for trial-based economic evaluations. For example, performance measures that aim to assess the extent to which the probability of cost-effectiveness is correctly estimated for various willingness-to-pay values. Further research into this area is warranted.

Bivariate model

Costs and effects are typically correlated. That is, patients with poor health outcomes might require more intensive treatments leading to higher costs, or, in contrast, patients with better health outcomes may have received more intensive treatment and thus have higher (treatment) costs. As a consequence, the correlation between cost and effects can bias cost-effectiveness estimates and is, therefore, ideally considered in a trial-based economic evaluation. In *Chapters 8* and *9* we explained the importance of considering the correlation between costs and effects in trial-based economic evaluations. Moreover, in a post hoc analysis of *Chapter 7,* we assessed whether specifying a bivariate

LLM according to the suggestion of Faria et al.[30] would perform better than a non-bivariate LLM. That is, costs and utility values at each time point were regressed upon the various covariates in the model simultaneously after rescaling their values to the same scale and adding a random intercept for the outcome. However, even though we did find that such a joint estimation of costs and QALY slightly improved the models' empirical bias, their coverage rates were found to be highly sensitive to an incorrect rescaling of costs and utility values, which makes them hard to apply in practice. In *Chapter 7*, we also found a non-bivariate MI-LLM to perform equally well as MI-SUR, which is bivariate in nature. This is in line with the results of Mutubuki et al.[48] who found accounting for the correlation between costs and effects not to have a large impact on cost and QALY estimates in two empirical datasets, nor on the statistical uncertainty surrounding both outcomes. Further research is needed to assess the relative performance of MI-LLM and MI-SUR versus Bayesian joint longitudinal models,[49] because such a modelling approach would enable the joint estimation of costs and effects, while also allowing the use of different distributions for both outcomes (e.g., Gamma for costs and Beta for QALY) and incorporating the longitudinal nature of the data.

Implications for research practice

For research practice, a number of implications can be formulated based on the studies presented in the thesis.

- *Chapters 2* and *3* indicate that caution is needed when using crosswalks for estimating utility values as they may significantly impact cost-utility outcomes, particularly in situations where the treatment effect size is small and in countries that were not included in the development of the crosswalks (i.e., all countries except Denmark, England, Italy, the Netherlands, Poland, and Scotland). For now, when EQ-5D value sets are not available, researchers and decision-makers should be careful in using crosswalks, especially when the country for which they perform their analysis is not included in the development of the crosswalk.
- *Chapter 4* indicates that country-specific value sets should be used in cost-utility analyses to account for the fact that health state preferences can be significantly affected by sociocultural differences.
- C*hapter 5* indicates that regression-based prediction models can be used to predict EQ-5D-3L utility values from ODI responses for use in trial-based economic evaluations among low back pain patients. However, they should not be used to predict HR-QoL for individual low back pain patients. *Chapter 6* further shows that response mapping approaches did not have an added value compared to the regression-based models of *chapter 5* when predicting EQ-5D-3L utility values from the ODI.
- Based on the results of *Chapter 7,* we recommend researchers multiply impute missing cost and effect observations first before using LLM when analysing longitudinal trial-based economic evaluation data. Alternatively, researchers can opt for a combination of MI and SUR as it has similar performance to MI combined with LLM, while it additionally accounts for the inherent correlation between costs and effects, and it is computationally faster.

Implications for further research

Various recommendations for further research have been provided in *Chapters 2* to *9*. Our most important recommendations include:

- *Chapters 2* to *6* show there is a need to develop performance measures that are specifically developed for trial-based economic evaluations. For example, performance measures that aim to assess the extent to which the probability of cost-effectiveness is correctly estimated for various willingness-to-pay values.
- To explore whether the finding of *Chapters 5* and *6* that response mapping approaches are not better than regression models is generalizable to other countries and populations (e.g., those with being with better health states) and the newest EQ-5D instrument (i.e., the EQ-5D-5L).
- As shown in *Chapter 7*, further research should assess the relative performance of MI-LLM and MI-SUR versus Bayesian joint longitudinal models because such a modelling approach would enable the joint estimation of costs and effects, while allowing the use of different distributions for both outcomes and incorporating the longitudinal nature of the data.

Conclusion

This thesis aimed to provide answers to methodological questions related commonly encountered in trial-based cost-effectiveness analyses. Findings suggest that the use of crosswalks and EQ-5D value sets can impact cost-utility outcomes. Country-specific EQ-5D value sets should be used in cost-utility analyses as population preferences considerably differ between countries. For the purpose of a cost-utility analysis, regression-based prediction models seem to be preferred over response mapping approaches to predict EQ-5D-3L utility values from the ODI when EQ-5D data were not collected but not to predict individual values. Moreover, MI is required before performing LLM for cost-effectiveness analysis to ensure that missing cost data is validly handled. All in all, the results of this thesis may have added value to improve the methodological quality of trial-based economic evaluations.

References

1. Emanuel EJ, Persad G, Upshur R, *et al.* Fair Allocation of Scarce Medical Resources in the Time of Covid-19. *New England Journal of Medicine* 2020; **382**: 2049–55.
2. Alhalaseh YN, Elshabrawy HA, Erashdi M, Shahait M, Abu-Humdan AM, Al-Hussaini M. Allocation of the "Already" Limited Medical Resources Amid the COVID-19 Pandemic, an Iterative Ethical Encounter Including Suggested Solutions From a Real Life Encounter. *Frontiers in Medicine* 2021; **7**. https://www.frontiersin.org/article/10.3389/fmed.2020.616277 (accessed May 12, 2022).
3. Muche-Borowski C, Abiry D, Wagner H-O, *et al.* Protection against the overuse and underuse of health care – methodological considerations for establishing prioritization criteria and recommendations in general practice. *BMC Health Serv Res* 2018; **18**. DOI:10.1186/s12913-018-3569-9.
4. Drummond MF, McGuire A. Economic Evaluation in Health Care: Merging Theory with Practice. Oxford University Press, 2001.
5. Tackling Wasteful Spending on Health | en | OECD. https://www.oecd.org/health/tackling-wasteful-spending-on-health-9789264266414-en.htm (accessed May 12, 2022).
6. Warner MA. Stop Doing Needless Things! Saving Healthcare Resources During COVID-19 and Beyond. *J GEN INTERN MED* 2020; **35**: 2186–8.
7. Drummond MF, Sculpher MJ, Torranc GW. Methods for the economic evaluation of health care programmes, 3rd ed. Oxford: Oxford University Press, 2005.
8. NWO. Netherlands Code of Conduct for Research Integrity | NWO. 2018. https://www.nwo.nl/en/netherlands-code-conduct-research-integrity (accessed May 23, 2022).
9. van Hout B, Janssen MF, Feng Y-S, *et al.* Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets. *Value in Health* 2012; **15**: 708–15.
10. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health- related quality of life: An EQ- 5D- 5L value set for England. *Health Econ* 2018; **27**: 7–22.
11. M. Versteegh M, M. Vermeulen K, M. A. A. Evers S, de Wit GA, Prenger R, A. Stolk E. Dutch Tariff for the Five-Level Version of EQ-5D. *Value in Health* 2016; **19**: 343–52.
12. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. *Medical Care* 2017; **55**: e51.
13. Mulhern B, Feng Y, Shah K, *et al.* Comparing the UK EQ-5D-3L and English EQ-5D-5L Value Sets. *PharmacoEconomics* 2018; **36**: 699–713.
14. Camacho EM, Shields G, Lovell K, Coventry PA, Morrison AP, Davies LM. A (five-) level playing field for mental health conditions?: exploratory analysis of EQ-5D-5L-derived utility values. *Qual Life Res* 2018; **27**: 717–24.
15. Yang F, Devlin N, Luo N. Cost-Utility Analysis Using EQ-5D-5L Data: Does How the Utilities Are Derived Matter? *Value Health* 2019; **22**: 45–9.
16. van Hout BA, Shaw JW. Mapping EQ-5D-3L to EQ-5D-5L. *Value in Health* 2021; published online May 18. DOI:10.1016/j.jval.2021.03.009.
17. Norman R, Cronin P, Viney R, King M, Street D, Ratcliffe J. International Comparisons in Valuing EQ-5D Health States: A Review and Analysis. *Value in Health* 2009; **12**: 1194–200.
18. Badia X, Roset M, Herdman M, Kind P. A Comparison of United Kingdom and Spanish General Population Time Trade-off Values for EQ-5D Health States. *Medical decision making* 2001; **21**: 7–16.
19. Johnson JA, Luo N, Shaw JW, Kind P, Coons SJ. Valuations of EQ-5D Health States: Are the United States and United Kingdom Different? *Medical Care* 2005; **43**: 221–8.

20. Lien K, Tam VC, Ko YJ, Mittmann N, Cheung MC, Chan KKW. Impact of country-specific EQ-5D-3L tariffs on the economic value of systemic therapies used in the treatment of metastatic pancreatic cancer. *Curr Oncol* 2015; **22**: 443.

21. Karlsson JA, Nilsson J-A, Neovius M, *et al.* National EQ-5D tariffs and quality-adjusted life-year estimation: comparison of UK, US and Danish utilities in south Swedish rheumatoid arthritis patients. *Annals of the Rheumatic Diseases* 2011; **70**: 2163–6.

22. Kiadaliri AA, Eliasson B, Gerdtham U-G. Does the choice of EQ-5D tariff matter? A comparison of the Swedish EQ-5D-3L index score with UK, US, Germany and Denmark among type 2 diabetes patients. *Health Qual Life Outcomes* 2015; **13**: 145.

23. van Hooff ML, Spruit M, Fairbank JCT, van Limbeek J, Jacobs WCH. The Oswestry Disability Index (Version 2.1a): Validation of a Dutch Language Version. *Spine* 2015; **40**: E83.

24. Carreon LY, Bratcher KR, Das N, Nienhuis JB, Glassman SD. Estimating EQ-5D Values From the Oswestry Disability Index and Numeric Rating Scales for Back and Leg Pain. *Spine* 2014; **39**: 678–82.

25. Fayers PM, Hays RD. Should Linking Replace Regression When Mapping from Profile-Based Measures to Preference-Based Measures? *Value in Health* 2014; **17**: 261–5.

26. Dorans NJ. Linking scores from multiple health outcome instruments. *Qual Life Res* 2007; **16**: 85–94.

27. Wailoo AJ, Hernandez-Alava M, Manca A, *et al.* Mapping to Estimate Health-State Utility from Non–Preference-Based Outcome Measures: An ISPOR Good Practices for Outcomes Research Task Force Report. *Value in Health* 2017; **20**: 18–27.

28. Dakin H. Review of studies mapping from quality of life or clinical measures to EQ-5D: an online database. *Health and Quality of Life Outcomes* 2013; **11**: 151.

29. Thompson NR, Lapin BR, Katzan IL. Mapping PROMIS Global Health Items to EuroQol (EQ-5D) Utility Scores Using Linear and Equipercentile Equating. *Pharmacoeconomics* 2017; **35**: 1167–76.

30. Faria R, Gomes M, Epstein D, White IR. A Guide to Handling Missing Data in Cost-Effectiveness Analysis Conducted Within Randomised Controlled Trials. *Pharmacoeconomics* 2014; **32**: 1157.

31. Gabrio A, Mason AJ, Baio G. Handling Missing Data in Within-Trial Cost-Effectiveness Analysis: A Review with Future Recommendations. *Pharmacoecon Open* 2017; **1**: 79–97.

32. Gabrio A, Plumpton C, Banerjee S, Leurent B. Linear mixed models to handle missing at random data in trial- based economic evaluations. *Health Economics* 2022; : hec.4510.

33. Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York, NY, USA: John Wiley & Sons, Inc., 2014.

34. Twisk J, de Boer M, de Vente W, Heymans M. Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of Clinical Epidemiology* 2013; **66**: 1022–8.

35. Peters SAE, Bots ML, Ruijter HM den, *et al.* Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models. *J CLIN EPIDEMIOL* 2012; **65**: 686–95.

36. Dong Y, Peng C-YJ. Principled missing data methods for researchers. *Springerplus* 2013; **2**: 222.

37. Little RJA. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics* 1988; **6**: 287–96.

38. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology* 2014; **14**: 75.

39. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; **38**: 2074–102.

40. Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* 2010; **21**: 128–38.

41. O'Kelly M, Anisimov V, Campbell C, Hamilton S. Proposed best practice for projects that involve modelling and simulation. *Pharmaceutical Statistics* 2017; **16**: 107–13.

42. Longworth L, Rowen D. Mapping to Obtain EQ-5D Utility Values for Use in NICE Health Technology Assessments. *Value in Health* 2013; **16**: 202–10.

43. Dakin H. Review of studies mapping from quality of life or clinical measures to EQ-5D: an online database. *Health and Quality of Life Outcomes* 2013; **11**: 151.

44. Dakin H, Abel L, Burns R, Yang Y. Review and critical appraisal of studies mapping from quality of life or clinical measures to EQ-5D: an online database and application of the MAPS statement. *Health and Quality of Life Outcomes* 2018; **16**: 31.

45. Petrou S, Rivero-Arias O, Dakin H, *et al*. The MAPS Reporting Statement for Studies Mapping onto Generic Preference-Based Outcome Measures: Explanation and Elaboration. *PharmacoEconomics* 2015; **33**: 993–1011.

46. Sculpher M, Claxton K, Akehurst R, Smith PC, Ginnelly L. It's just evaluation for decision-making: recent developments in, and challenges for, cost-effectiveness research. 2005: 8–41.

47. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics* 1999; **18**: 341–64.

48. Mutubuki EN, El Alili M, Bosmans JE, *et al*. The statistical approach in trial-based economic evaluations matters: get your statistics together! *BMC Health Serv Res* 2021; **21**: 475.

49. Gabrio A, Baio G, Manca A. Bayesian Statistical Economic Evaluation Methods for Health Technology Assessment. In: Oxford Research Encyclopedia of Economics and Finance. Oxford University Press, 2019. DOI:10.1093/acrefore/9780190625979.013.451.

CHAPTER 11

Summary

Background

Healthcare decision-makers around the world are facing tough decisions regarding the allocation of healthcare resources in the context of restrained supplies and escalated costs. To inform such allocation decisions, researchers are called upon to demonstrate not only the effectiveness of healthcare interventions but also their cost-effectiveness. Trial-based economic evaluations seek to provide this information by relating the difference in costs between healthcare interventions to the difference in effects. In this type of study, health-related quality of life (HRQoL) is often used as a metric of health effects. A prerequisite for using results of trial-based economic evaluations in decision-making is that they are valid and reliable, i.e., "*scrupulous*". Using less-than-optimal methods for estimating HRQoL and the statistical analysis of cost-effectiveness data can lead to biased conclusions and thus potentially a waste of healthcare resources. Despite this knowledge, the literature indicates that the methodological quality of published trial-based economic evaluations is generally suboptimal. This PhD project aimed, therefore, to address some of the gaps in the validity of HRQoL measures and statistical challenges in applied economic evaluations. This was done by answering research questions that emerged from the analysis of trial-based economic evaluation data and by providing step-by-step guidance on how to conduct and interpret trial-based economic evaluations based on statistically sound methods available in the literature.

Part 1 Methodological studies

Chapter 2 compared the EQ-5D-5L crosswalks and the 5L value sets for England, the Netherlands, and Spain and explored the implication of using one or the other for healthcare decision-making. This was done by using data from two randomized controlled trials on depression and diabetes. Both utility scoring methods were described by country and compared in terms of utility values distributions, mean values, and QALYs that were calculated using the area-under-the-curve method. Subsequently, cost-effectiveness analyses were performed to calculate ICERs. The uncertainty around ICERs was estimated using bootstrapping and probabilities of cost-effectiveness were graphically shown in cost-effectiveness acceptability curves. For all countries investigated, utility value distributions differed between the EQ-5D-5L crosswalk and the 5L value set. In both case studies, mean utility values were lower for the EQ-5D-5L crosswalk compared with the 5L value set in England and Spain, but higher in the Netherlands. However, these differences in utility values between the EQ-5D scoring methods did not translate into relevant differences in incremental QALYs, ICERs, and the interventions' probability of cost-effectiveness in the investigated countries. Results, therefore, suggested that EQ-5D-5L crosswalks and 5L value sets could be used interchangeably in patients affected by mild or moderate conditions. However, findings also indicated the need for further research to establish whether these results are generalizable to trial-based economic evaluations among severely ill patients.

Chapter 3 sought to further investigate whether the results of *Chapter 2* could be generalizable to other patient conditions in a simulation study. Trial-based economic evaluation data were then simulated for different conditions (depression, low back pain, osteoarthritis, cancer), severity levels (mild, moderate, severe), and effect sizes (small, medium, large). For all 36 scenarios, utility values were calculated using 3L and 5L value sets and crosswalks (3L to 5L and 5L to 3L crosswalks) for the Netherlands, the United States, and Japan. Utility values, QALYs, incremental QALYs, ICERs, and probabilities of cost-effectiveness obtained from values sets and crosswalks were subsequently compared. Differences between value sets and crosswalks ranged from -0.33 to 0.13 for utility values, from -0.18 to 0.13 for QALYs, and from -0.01 to 0.08 for incremental QALYs, resulting in different ICERs. For small effect sizes, at a willingness-to-pay of €20,000/QALY, the largest difference in probability of cost-effectiveness was found for moderate cancer between the Japanese 5L value set and 5L to 3L crosswalk (difference=0.63). For medium effect sizes, the largest difference was found for mild cancer between the Japanese 3L value set and 3L to 5L crosswalk (difference=0.06). For large effect sizes, the largest difference was found for mild osteoarthritis between the Japanese 3L value set and 3L to 5L crosswalk (difference=0.08). Differently from *Chapter 2,* the findings of *Chapter 3* indicated that reimbursement decisions may change in situations with small effect sizes and countries that were not included in the development of the crosswalks. Therefore, when EQ-5D value sets are not available, researchers and decision-makers should be aware that the use of crosswalks is likely to impact decisions.

Chapter 4 assessed the impact of EQ-5D country-specific value sets on cost-utility outcomes by using data from two randomized controlled trials on low back pain and depression. 3L value sets for 16 countries were used and a nonparametric crosswalk was employed for each tariff to obtain the likely 5L values. Differences in QALYs between countries were tested using paired t-tests, with the EQ-5D United Kingdom value set as the reference. Cost-utility outcomes were estimated for both studies and both EQ-5D-3L versions, including differences in QALYs and cost-effectiveness acceptability curves. For the 3L, QALYs ranged between 0.650 (Taiwan) and 0.892 (United States) in the LBP study and between 0.619 (Taiwan) and 0.879 (United States) in the depression study. In both studies, most country-specific QALY estimates differed statistically significantly from that of the United Kingdom. ICERs ranged between €2044/QALY (Taiwan) and €5897/QALY (Zimbabwe) in the low back pain study and between €38,287/QALY (Singapore) and €96,550/QALY (Japan) in the depression study. At the NICE threshold of €23,300/QALY (≈£20,000/QALY), the intervention's probability of being cost-effective versus control ranged between 0.751 (Zimbabwe) and 0.952 (Taiwan) and between 0.230 (Canada) and 0.396 (Singapore) in the LBP study and depression study, respectively. Similar results were found for the mapped 5L, with extensive differences in ICERs and moderate differences in the probability of cost-effectiveness. Results indicated that the use of different EQ-5D country-specific value sets impacts cost-utility outcomes. Therefore, to account for the fact that health state preferences are affected by sociocultural differences, relevant country-specific value sets should be used.

Chapter 5 investigated whether EQ-5D-3L utility values could be predicted from the ODI in cost-effectiveness analysis to compare interventions targeting patients with low back pain patients. Regression models were developed in a random sample of 70% of 18,692 patients with low back pain whose data was previously collected on both EQ-5D-3L and ODI simultaneously. EQ-5D-3L utility values were estimated using their ODI scores as independent variables in six different models, namely: 1) Ordinary Least Squares (OLS) regression, with total ODI score, 2) OLS, with ODI item scores as continuous variables, 3) OLS, with ODI item scores as ordinal variables, 4) Tobit model, with total ODI score, 5) Tobit model, with ODI item scores as continuous variables, 6) Tobit model, with ODI item scores as ordinal variables. The models' performance was assessed in the remaining 30% of the sample using explained variance ($R^{2)}$ and Root Mean Squared Error (RMSE). The potential impact of using predicted instead of observed EQ-5D-3L utility values on cost-effectiveness outcomes was evaluated in two empirical cost-effectiveness analyses. All models had a relatively similar R2 (range: 45-52%) and RMSE (range: 0.21-0.22). The two best-performing models produced similar probabilities of cost-effectiveness for a range of willingness-to-pay (WTP) values compared to those based on the observed EQ-5D-3L values. For example, the difference in probabilities ranged from 2% to 5% at a WTP of 50,000 €/QALY gained. Results of *Chapter 5* suggested that the ODI can be used to predict EQ-5D utility values for use in cost-effectiveness analysis. However, the models are not suitable for estimating utility values for individual patients. Findings indicated that further research is needed to validate the models in other empirical datasets on low back pain interventions. Additional investigation was also suggested to compare the performance of predictive modeling with that of other mapping approaches for estimating utility values.

Chapter 6 built on the findings of *Chapter 5* by developing and validating approaches for mapping ODI responses to EQ-5D-3L utility values and evaluating the impact of using mapped utility values on cost-utility results compared to published regression models. Three response mapping approaches were developed using the 70% sample of low back pain patients used in *Chapter 5*: 1) non-parametric approach (Non-p), 2) non-parametric approach excluding logical inconsistencies (Non-peLI), and 3) ordinal logistic regression (OLR). Performance was assessed in the remaining 30% using R-square (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). To evaluate whether MAEs and their 95% limits of agreement (95%LA) were clinically relevant, a minimally clinically important difference (MCID) of 0.074 was used. Probabilities of cost-effectiveness estimated using observed and mapped utility values were compared in two economic evaluations. The Non-p performed best (R^2=0.43; RMSE=0.22; MAE=0.03; 95%LA=-0.40;0.47) compared to the Non-peLI (R^2=0.07; RMSE=0.29; MAE=-0.15; 95%LA=-0.63;0.34), and ORL (R^2=0.22; RMSE=0.26; MAE=0.02; 95%LA=-0.49;0.53). MAEs were lower than the MCID for the Non-p and OLR, but not for the Non-peLI. Differences in probabilities of cost-effectiveness ranged from 1-4% (Non-p), 0.1-9% (Non-peLI), and 0.1-20% (OLR). Results suggested that the developed response mapping approaches are not valid for estimating individual patients' EQ-5D-3L utility values, and – depending on the approach – may considerably impact cost-utility results. Response mapping approaches did not perform better than previously published regression-based models and are, therefore, not recommended for use in economic evaluations.

Chapter 7 investigated whether MI is required prior to LLM when analysing longitudinal cost and effect data as the literature suggests otherwise when analysing effect data. Two-thousand complete datasets were simulated containing five-time points. Incomplete datasets were generated with 10%, 25%, and 50% missing data in follow-up costs and effects, assuming a Missing At Random (MAR) mechanism. Six different strategies were compared using empirical bias (EB), root-mean-squared error (RMSE), and coverage rate (CR). These strategies were: LLM alone (LLM) and MI with LLM

(MI-LLM), and, as reference strategies, mean imputation with LLM (M-LLM), seemingly unrelated regression alone (SUR-CCA), MI with SUR (MI-SUR), and mean imputation with SUR (M-SUR). For costs and effects, LLM, MI-LLM, and MI-SUR performed better than M-LLM, SUR-CCA, and M-SUR, with smaller EBs and RMSEs as well as CRs closers to nominal levels. However, even though LLM, MI-LLM, and MI-SUR performed equally well for effects, MI-LLM and MI-SUR were found to perform better than LLM for costs at 10% and 25% missing data. At 50% missing data, all strategies resulted in relatively high EBs and RMSEs for costs. Based on these findings, it is advisable to multiply impute missing values before LLM for analysing trial-based economic evaluation data. One might also opt for a combination of MI and SUR. MI-SUR is computationally more efficient than MI-LLM, but it does not allow for the estimation of average intervention effects over time while having the advantage of accounting for the correlation between costs and effects.

Part 2 Tutorials

Chapter 8 provided step-by-step guidance on how to combine appropriate statistical methods available in the literature for handling methodological challenges encountered in trial-based economic evaluations, namely: 1) missing data, 2) correlated costs and effects, 3) baseline imbalances, and 4) skewness of costs and/or effects. This was done by using a ready-to-use R script, including annotated code providing an explanation of the theoretical background of each method. Although the provided R script is expected to be suitable for use in most trial-based cost-effectiveness analyses, there may be other specific methodological challenges, e.g., using a linear mixed model as an analysis model, or analysing multi-arm trials for which were presented possible ways to adapt the provided annotated R code. *Chapter 8* also discussed the limitations of the approach chosen in this tutorial and pointed out topics for further research, e.g., the use of Bayesian models as compared to usually used frequentist ones.

Chapter 9 aimed to further illustrate and explain the basic concepts when designing, analysing, and interpreting trial-based economic evaluations by using a case study assessing the cost-effectiveness of exercise therapy compared to standard advice in patients with musculoskeletal disorders. *Chapter 9* was a project conducted in collaboration with fellow colleagues from Brazil who identified the need to support Brazilian healthcare decision-makers with the interpretation of economic evaluations in the field of musculoskeletal disorders and to aid physiotherapists with translating the results of such studies into clinical practice.

Discussion

In *Chapter 10*, the main findings of this thesis were discussed and interpreted based on the existing literature as well as methodological recommendations were additionally provided. In conclusion, researchers and decision-makers need to be aware that the use of crosswalks is likely to impact cost-utility results to such an extent that this may affect reimbursement decisions. This is particularly warranted in situations where treatment effect sizes are relatively small and in countries that were not included in the development of the EQ-5D crosswalk. Additionally, to account for the fact that health state preferences are affected by sociocultural differences, the most appropriate EQ-5D scoring method is the available country-specific value set. Another relevant conclusion is that in the absence of the EQ-5D in studies investigating interventions on low back pain, the condition-specific measure ODI might be used to predict utility values for use in cost-effectiveness analysis but not to estimate HRQoL for individual patients. The latter is due to the low fit of regression models and poor agreement between mapped and observed utility values. The former may be explained by the fact that the bias surrounding the predicted utility values is likely to be similar in the intervention and control groups, thereby not affecting incremental QALYs. In the context of trial-based economic evaluations, mapping approaches such as regression models are preferred over response mapping ones as the former models outperformed the latter ones and are more likely to be known by health economists making them easier to implement. Researchers should also acknowledge that mapping approaches are interim solutions and the most appropriate course of action is to plan beforehand the inclusion of generic HRQoL measures in the design of clinical trials when the aim is to perform a cost-effectiveness analysis. It should be also noted that when using LLM as an analysis model for cost-effectiveness analysis and missing data is related to the observed variables, researchers are advised to multiply impute missing values first as MI is more likely to produce unbiased cost estimates while accounting for the uncertainty around imputed values. To facilitate researchers in applying the methods presented in this thesis software codes and tutorials were provided in line with good practices in research namely scrupulousness and transparency.

CHAPTER 12

Samenvatting

Achtergrond

Beleidsmakers wereldwijd hebben steeds vaker te maken met vraagstukken omtrent de verdeling van schaarse financiële middelen in de zorg. Dit wordt steeds complexer door een groter wordend tekort aan financiële middelen in de zorg enerzijds en een stijging in zorgkosten anderzijds. Om dergelijke beslissingen te kunnen maken, worden onderzoekers steeds vaker verzocht niet alleen de effectiviteit van interventies aan te tonen, maar ook de kosteneffectiviteit. Empirische economische evaluaties kunnen aan dit verzoek voldoen door het verschil in kosten tussen interventies te relateren aan het daarbij behorende verschil in gezondheidseffecten. Dergelijke gezondheidseffecten worden in dit soort studies vaak uitgedrukt in gezondheids-specifieke kwaliteit van leven (health-related quality of life; [HRQoL]).

Een belangrijke voorwaarde voor het gebruik van de resultaten van empirische economische evaluaties tijdens besluitvormingsprocessen is dat deze resultaten valide en betrouwbaar zijn. Het gebruik van suboptimale methoden voor het schatten van HRQoL en/of het gebruik van suboptimale statistische methoden kan leiden tot niet valide en onbetrouwbare resultaten en conclusies, en daarmee ook tot verkeerde beslissingen, wat weer ten gevolg kan hebben dat schaarse financiële middelen worden verspild. Ondanks het feit dat dit een bekend fenomeen is, laat de methodologische kwaliteit van gepubliceerde empirische economische evaluaties vaak te wensen over. Dit proefschrift heeft daarom tot doel om een aantal kennishiaten op te vullen betreffende het meten en waarderen van kwaliteit van leven en de statistische analyse van empirische economische evaluaties. Deze hiaten betreffen: 1) de invloed van het gebruik van crosswalks op de besluitvorming in de zorg, 2) het schatten van HRQoL op basis van een ziekte-specifieke patiënt-gerapporteerde uitkomstmaat (patient-reported outcome measure; PROM), en 3) het omgaan met missende data in empirische economische evaluaties. Bovendien bevat dit proefschrift twee tutorial artikelen over hoe een empirische economische evaluatie te ontwerpen, uit te voeren, te analyseren en te interpreteren met als doel bij te dragen aan het gebruik van optimale en valide methoden in toekomstige empirische economische evaluaties.

Deel 1 Methodologische studies

In *hoofdstuk 2* werden de EQ-5D-5L crosswalks en de 5L waarderingsets voor Engeland, Nederland en Spanje met elkaar vergeleken. Daarnaast werd ook de eventuele invloed van het gebruik van één van beide waarderingsmethoden (crosswalks vs. waarderingsets) op de daarmee samenhangende besluitvorming onderzocht. Hiervoor werd data van twee gerandomiseerde klinische onderzoeken gebruikt, waarvan de ene een interventie betrof op het gebied van diabetes en de andere een interventie op het gebied van depressie. Per land, werden de geschatte utiliteiten en QALY's vergeleken tussen beide waarderingsmethoden. Vervolgens werden volledige empirische economische evaluaties uitgevoerd voor de twee gerandomiseerde klinische onderzoeken, gebruikmakend van beide waarderingsmethoden voor Engeland, Nederland en Spanje. Incrementele kosteneffectiviteitsratio's (ICER's) werden berekend en de onzekerheid rondom de

ICER's werd geschat door middel van de non-parametrische bootstrap. De kans op kosteneffectiviteit werd daarnaast weergegeven in zogenaamde "cost-effectiveness acceptability curves" (CEACs). De resultaten lieten zien dat de verdeling van de geschatte utiliteiten verschilden tussen de waarderingsmethoden. Daarnaast lieten de resultaten zien dat utiliteiten over het algemeen lager waren wanneer er gebruik werd gemaakt van de EQ-5D-5L crosswalk in plaats van de 5L waarderingset. Echter, deze verschillen in utiliteiten vertaalden zich niet in relevante verschillen in incrementele QALY's, ICER's en de kans op kosteneffectiviteit voor de drie landen. Dit betekent dat op basis van de resultaten gesteld kan worden dat zowel de EQ-5D-5L crosswalk als de 5L waarderingset gebruikt kan worden in empirische economische evaluaties van ziektebeelden met een milde tot gemiddelde ziektelast in Engeland, Nederland en Spanje. Toekomstig onderzoek is nodig om te evalueren of deze resultaten te generaliseren zijn naar ernstige ziektebeelden.

In *hoofdstuk 3* werd uitgezocht of de resultaten uit *hoofdstuk 2* te generaliseren zijn naar ziekten met een ernstigere ziektelast. Dit werd gedaan middels een simulatie studie. Kosten en effect data werd gesimuleerd voor vier verschillende ziektebeelden (depressie, lage rugpijn, osteoartritis en kanker), drie maten van de ernst van ziekte (mild, gemiddeld en ernstig) en drie effectgroottes (klein, gemiddeld en groot). Voor alle 36 scenario's werden utiliteiten, incremental QALY's en ICER's geschat op basis van de 3L en 5L waarderingsets en crosswalks (3L naar 5L en 5L naar 3L crosswalks) voor Nederland, de Verenigde Staten en Japan. De resultaten lieten zien dat verschillen tussen waarderingssets en crosswalks uiteenliepen van -0,33 tot 0,13 voor utiliteiten, van -0,18 tot 0,13 voor QALY's en van -0,01 tot 0,08 voor incrementele QALY's. Dit resulteerde tevens in verschillende ICER's. Voor kleine effectgroottes, en bij een referentiewaarde van 20,000/QALY, werd het grootste verschil gevonden in de kans op kosteneffectiviteit bij een milde vorm van kanker tussen de Japanse 5L waarderingset en de 5L naar 3L crosswalks (verschil=0,63). Voor gemiddelde effectgroottes, werd het grootste verschil gevonden in de kans op kosteneffectiviteit bij een milde vorm van kanker tussen de Japanse 3L waarderingset en de 3L naar 5L crosswalks (verschil=0,06). Voor grote effectgroottes, werd het grootste verschil gevonden in de kans op kosteneffectiviteit bij een milde vorm van osteoartritis tussen de Japanse 3L waarderingset en de 3L naar 5L crosswalks (verschil=0,08). In tegenstelling tot de resultaten van hoofdstuk 2, lijkt het erop dat vergoedingsbesluiten wel degelijk kunnen worden beïnvloed door de keuze tussen het gebruik van een waarderingsset of een crosswalk, en met name in landen die niet geen deel uitmaakten van de studie die ten grondslag lag aan de crosswalks (bijvoorbeeld Japan en de Verenigde Staten). Wanneer EQ-5D waarderingsets niet beschikbaar zijn voor een bepaald land, is het daarom van belang dat beleidsmakers zich er bewust van zijn dat het gebruik van een crosswalks weldegelijk invloed kan hebben op de resultaten van een economische evaluatie en daarmee op het bijbehorende besluitvormingsproces in de zorg.

In *hoofdstuk 4* werd de invloed van het gebruik van verschillende land-specifieke EQ-5D waarderingsets op de resultaten van economische evaluaties geëvalueerd. Hiervoor werd gebruik gemaakt van data uit twee gerandomiseerde klinische onderzoeken op het gebied van lage rugpijn en depressie. In totaal werden 3L waarderingsets van 16 verschillende landen vergeleken en een non-parametrische crosswalk was gebruikt om ook inzicht te krijgen in de eventuele verschillen tussen land-specifieke waarderingssets voor de 5L. Verschillen in QALY's tussen landen werden statistisch getoetst door middel van een paarsgewijze t-toets, met de Britse EQ-5D waarderingset als referentie. Economische evaluatie resultaten, o.a. incrementele QALY's en CEAC's, werden geschat voor beide studies gebruikmakend van de verschillende land-specifieke waarderingsets en voor beide EQ-5D versies (3L en 5L). Voor de 3L liepen de QALY's uiteen van 0,650 (Taiwan) tot 0,892 (Verenigde Staten) in de lage rugpijn studie en tussen 0,619 (Taiwan) en 0,879 (Verenigde Staten) in de depressie studie. ICER's liepen uiteen van 2.044/QALY (Taiwan) tot 5.897/QALY (Zimbabwe) in de lage rugpijn studie en van 38.287/QALY (Singapore) tot 96550/QALY (Japan) in de depressie studie. Bij een referentiewaarde van 23.300/QALY (20000/QALY), zoals gehanteerd door NICE, liep de kans op kosteneffectiviteit uiteen van 0,751 (Zimbabwe) tot 0,952 (Taiwan) in de lage rugpijn studie en van 0,230 (Canada) tot 0,396 (Singapore) in de depressie studie. Vergelijkbare resultaten werden gevonden voor de geschatte 5L utiliteiten (op basis van mapping), met grote verschillen in ICER's en wat kleinere verschillen in de kans op kosteneffectiviteit. Resultaten uit dit hoofdstuk suggereren dat gebruik van verschillende land-specifieke EQ-5D waarderingsets invloed kan hebben op de resultaten van een economische evaluatie. Om rekening te houden met het feit dat voorkeuren van mensen voor verschillende gezondheidstoestanden worden beïnvloed door socioculturele verschillen, is het daarom van belang dat – indien beschikbaar – de EQ-5D waarderingsset wordt gebruikt van het land waar een bepaalde studie is verricht.

In *hoofdstuk 5* werd onderzocht of EQ-5D-3L utiliteiten kunnen worden voorspeld op basis van de ODI en of deze op valide wijze gebruikt kunnen worden in economische evaluaties van lage rugpijn interventies. Regressiemodellen werden ontwikkeld in een willekeurig selectie van 70% van de 18.692 patiënten met lage rugpijn waarvan zowel EQ-5D-3L als ODI data beschikbaar was. De EQ-5D-3L utiliteiten werden als afhankelijk variabele meegenomen en de ODI scores als onafhankelijke variabelen in zes verschillende modellen: 1) ordinary least squares (OLS) regressie met totale ODI score als onafhankelijke variabele, 2) OLS, met ODI item scores als continu variabelen, 3) OLS met ODI item scores als ordinale variabelen, 4) Tobit model met totale ODI score als onafhankelijke variabele, 5) Tobit model met ODI item score als continu variabelen en 6) Tobit model met ODI item scores als ordinale variabelen. "Model performance" werd geëvalueerd in de overige 30% van de patiënten middels de R-squared (R^2) en de Root Mean Squared Error (RMSE). De mogelijke impact van het gebruik van voorspelde in plaats van geobserveerde EQ-5D-3L utiliteiten op de resultaten van economische evaluaties werd onderzocht in twee case studies. De resultaten lieten zien dat alle modellen een relatief vergelijkbaar R^2 (range: 45-52%) en RMSE (range: 0,21-0,22) hadden. Hoewel deze R^2 waarden suggereren dat de ODI een groot deel van de variantie in EQ-5D-3L utiliteiten verklaarde, geven de hoge RMSE waarden aan dat er een relatief groot verschil zit tussen de voorspelde en geobserveerde EQ-5D-3L utiliteiten. Opmerkelijk was dat de modellen met de beste performance, namelijk model 2 en model 5, toch resulteerden in relatief vergelijkbare voorspelde kansen op kosteneffectiviteit bij verschillende referentiewaarden vergeleken met de geobserveerde waarden. Bijvoorbeeld, het verschil in de kans op kosteneffectiviteit liep slechts uiteen van 3% tot 5% bij een referentiewaarde van 50,000/QALY voor respectievelijk model 2 en model 5. Op basis van de huidige resultaten kan daarom gesteld worden dat de ODI gebruikt kan worden om utiliteiten te voorspellen in economische evaluaties van lage rugpijn interventies indien de EQ-5D niet is afgenomen. Echter, gezien de hoge RMSE waarden, zijn deze modellen niet geschikt om utiliteitswaarden te schatten van individuele patiënten. Verder onderzoek is nodig om de onderzochte modellen te valideren in andere empirische datasets. Bovendien is het van belang om de performance van de ontwikkelde regressiemodellen te vergelijken met die van mapping methoden.

In *hoofdstuk 6* werd verder gebouwd op de resultaten uit *hoofdstuk 5*. In dit hoofstuk werden verschillende mapping modellen ontwikkeld en gevalideerd voor het schatten van EQ-5D-3L utiliteiten op basis van de ODI. Daarnaast werden de ontwikkelde modellen vergeleken met reeds gepubliceerde regressiemodellen (zie *hoofdstuk 5*). Drie response mapping modellen werden ontwikkeld op basis van 70% van de beschikbare data: 1) non-parametrisch model (Non-p), 2) non-parametrisch model, exclusief logische inconsistenties (non-peLI) en een 3) ordinale logistische regressie model (ORL). "Model performance" werd geëvalueerd in de overige 30% van de data middels de R-squared (R^2), de Root Mean Squared Error (RMSE) en de Mean Absolute Error (MAE). Om te evalueren of de MAE's en de bijbehorende limits of agreement (95%LA) klinisch relevant waren, werd een minimally clinically important difference (MCID) van 0,074 aangehouden. De Non-p bleek de beste model performance te hebben (R^2=0,43; RMSE=0,22; MAE=0,03; 95%LA=-0,40;0,47) vergeleken met de NonpeLI (R^2=0,07; RMSE=0,29; MAE=-0,15; 95%LA=-0,63;0,34), en de ORL (R^2=0,22; RMSE=0,26; MAE=0,02; 95%LA=-0,49;0,53). MAE's waren kleiner dan de MCID voor de Non-p en de OLR, maar niet voor de Non-peLI. De verschillen in de kans op kosteneffectiviteit liepen uiteen van 1% tot 4% (Non-p), 0,1% tot 9% (Non-peLI) en 0,1% tot 20% (OLR). Op basis van de resultaten van *hoofdstuk 6* kan gesteld worden dat de ontwikkelde response mapping modellen niet valide zijn om EQ-5D-3L utiliteiten te schatten van individuele patiënten, en dat het gebruik van sommige van deze modellen ook een significante invloed kan hebben op de resultaten van een economische evaluatie. Daarnaast bleek de performance van de ontwikkelde response mapping modellen niet beter te zijn dan die van reeds gepubliceerde regressiemodellen. Daarmee kunnen de ontwikkelde response mapping modellen dan ook niet worden aanbevolen voor gebruik in economische evaluaties.

In *hoofdstuk 7* werd onderzocht of MI nodig voordat kosten en effect data longitudinaal geanalyseerd werd middels LLM in een empirische economische evaluatie. Tweeduizend complete datasets werden gesimuleerd met vijf verschillende tijdspunten. Daaruit werden incomplete datasets gemaakt met 10%, 25% en 50% missende data in follow-up kosten en effecten, op basis van een Missing At Random (MAR) mechanisme. De statistische performance van de zes verschillende strategieën werden met elkaar vergeleken middels empirical bias (EB), root mean squared error (RMSE) en coverage rate (CR). Deze strategieën waren: 1) LLM (LLM), 2) MI met LLM (MI-LLM), en als referentie strategieën, 3) mean imputatie met LLM (M-LLM), 4) complete-case analyse middels seemingly unrelated regression (SUR-CCA), 5) MI met SUR en 6) mean imputatie met SUR (M-SUR). Voor kosten en effecten, resulteerden LLM, MI-LLM en MI-SUR in een betere performance dan M-LLM, SUR-CCA en M-SUR. Echter, alhoewel de performance van LLM, MI-LLM en MI-SUR vergelijkbaar was voor effecten, resulteerden MI-LLM en MI-SUR in een betere performance dan LLM voor kosten bij 10% en 25% missende data. Bij 50% missende data resulteerden alle strategieën in relatief hoge EB's en RMSE's voor kosten. Op basis van deze resultaten is het aan te bevelen missende waarden eerst te imputeren middels MI gevolgd door een LLM indien men kosten en effect data longitudinaal wil analyseren in een empirische economische evaluatie. Er kan ook gekozen worden voor MI-SUR, echter, ook al is deze methode rekenkundig efficiënter dan MI-LLM, het kan geen overall interventie effect over tijd schatten. Voordeel blijft echter wel dat er een correctie kan plaatsvinden voor de eventuele correlatie tussen kosten en effecten.

Deel 2 Tutorials

Hoofdstuk 8 bevat stap-voor-stap aanbevelingen ten behoeve van het analyseren van economische evaluaties. Hierbij is rekening gehouden met vier statistische uitdagingen die veelvoudig voorkomen in dergelijke studies, namelijk: 1) missende data, 2) gecorreleerde kosten en effecten, 3) baseline verschillen en 4) scheefverdeeldheid van kosten en/of effecten. Daarnaast bevat dit hoofdstuk een *ready-to-use* R script, welke andere onderzoekers kunnen gebruiken om hun economische evaluatie data zo valide mogelijk te evalueren. Hoewel het R script gebruikt kan worden in studies die gekenmerkt worden door de vier hierboven beschreven statistische uitdagingen, zijn er ook andere uitdagingen waarmee onderzoekers te maken kunnen krijgen. Voorbeelden hiervan zijn het longitudinaal willen analyseren van kosten en effect data, en multi-armige klinische onderzoeken. Het R script kan relatief eenvoudig aangepast worden om ook met deze statistische uitdagingen om te gaan en hier zijn enkele suggesties voor gedaan.

Hoofdstuk 9 had als doel om de basisconcepten gerelateerd aan het ontwerp, de analyse en de interpretatie van empirische economische evaluaties uit te leggen en te illustreren aan de hand van een case studie. In deze case studie werd de kosteneffectiviteit geschat van een Pilates interventie voor patiënten met klachten aan het houdings- en bewegingsapparaat vergeleken met standaard zorg. *Hoofdstuk 9* is geschreven in nauwe samenwerking met collega's uit Brazilië die aangaven ondersteuning nodig te hebben bij de interpretatie van economische evaluaties en om fysiotherapeuten te ondersteunen bij het vertalen van kosteneffectiviteitsresultaten naar de klinische praktijk.

Discussie

In *hoofdstuk 10* werden de resultaten van deze thesis besproken en geïnterpreteerd op basis van bestaande literatuur. Daarnaast werd een aantal methodologische aanbevelingen gedaan. Onderzoekers werden er onder andere op gewezen dat zij op de hoogte dienden te zijn dat het gebruik van EQ-5D crosswalks in plaats van EQ-5D waarderingssets mogelijk invloed kan hebben op de resultaten van economische evaluaties en de en daarmee samenhangende vergoedingsbeslissingen. Dit is met name belangrijk in situaties waar er sprake is van kleine effectgroottes en als het gaat om landen die niet zijn meegenomen in de ontwikkeling van de EQ-5D crosswalk. Daarnaast kan – indien aanwezig - het beste gebruik worden gemaakt van de land-specifieke EQ-5D waarderingset van het land in kwestie. Bij een gebrek aan EQ-5D data kunnen deze geschat worden middels regressiemodellen op basis van de ziekte-specifieke uitkomstmaat ODI voor gebruik in economische evaluaties van lage rugpijn interventies, maar niet om de HRQoL van individuele patiënten te schatten. In een andere studie werd zelfs gevonden dat dergelijke regressiemodellen geprefereerd dienden te worden over mapping modellen. Onderzoekers moeten zich echter wel bewust zijn dat dergelijke modellen interim-oplossingen zijn en dat het altijd het beste is om de EQ-5D zelf bij patiënten af te nemen. Ook is het belangrijk dat wanneer LLM wordt gebruikt in een economische evaluaties om kosten en effect data longitudinaal te analyseren missende data idealiter eerst geïmputeerd wordt middels MI. Om onderzoekers te faciliteren in het toepassen van de gepresenteerde methoden bevat dit proefschrift twee tutorial artikelen, inclusief de benodigde software codes.

CHAPTER 13

Acknowledgments

I would like to start by saying that I am super grateful for having (had) a PhD journey, during which I believe I was able to develop to the best of my potential as a scientist thanks to the guidance and endless support of my supervisors **Judith Bosmans**, **Hanneke van Dongen**, and **Mohamed El Alili**. I have learned a lot from you through your valuable and insightful supervision throughout my PhD. **Judith**, I highly value your honest, fair, and timely feedback, which generally opened my eyes to my blind spots and helped me to sharpen my senses and to find better ways of tackling important issues. You have always been aware of my skills and potential, and you have truly encouraged me to become a good researcher as much as possible. You offered me countless opportunities to demonstrate my skills and abilities as a health economist and gave me the opportunity to work on a broad range of trial-based economic evaluation projects of varying health interventions. From these enriching experiences emerged several methodological questions that eventually formed the basis of my PhD thesis. Thank you for sharing your immense amount of knowledge with me and for believing in my potential to fulfil the required tasks. **Hanneke van Dongen**, I really appreciate you being there at each step of my learning process, bringing up your bright perspectives, enlightening me with your sharp way of reasoning in such a nice and cool way, giving me the heads up if needed, and pushing me to move forward while caring for my needs. **Hanneke** and **Mohamed**, you have encouraged me to *get my statistics together*! Also, your clever and straightforward way of doing research and the way you put results into perspective is outstanding and has motivated me throughout my PhD trajectory. I am confident to say that the three of you have provided me with the means to become a true scientist. I can't thank you enough for the empowering feeling that has flourished in me during your supervision.

A big thank you also goes to all my co-authors for their contributions to the work in this thesis. It was a pleasure to work with you and to learn from you! **Aureliano Finch**, I really enjoyed having you as a roommate at VU and I greatly appreciate your expert guidance in the health-related quality of life field, which made me aware of important research gaps and hence come up with new research ideas. **Sylvia Pellekooren**, it was enriching – and above all fun – to work with you on the "van Meenen middelen project", to perform analyses in R, to debug bugs, and to figure out ways to make our analysis codes more efficient. **Raymond Ostelo**, thank you for the insightful expert guidance in clinimetrics and low back pain as well as for the insights into research-life-related topics.

I am also grateful to the doctorate committee, **Elly Stolk**, **Jan van Busschbach**, **Ardine de Wit**, **Wendy Janssens**, and **Erik Koffijberg** for the time and effort they took to critically assess my thesis. You are definitely a source of inspiration for my work as a scientist and I look forward to our discussion about my thesis.

An important part of my "PhD life" was filled with wonderful colleagues in the Health Technology Assessment section of the VU. **Alejandra Guevara Morel** thank you for giving me the heads up for the practical PhD trajectory tasks and for making me aware of the fact that there was more than research; e.g., having fun in a day out with family and friends at the Efteling; taking time to do nice things to myself. Also, it has been super fun to work on your simulation study, and to code and debug analyses in R with you! **Anita Varga**, thank you for the very clever econometric-related tips (e.g., checking my equations) as well as your tips about work-life balance! **Elizabeth Mutubuki**, **Hana**

Broulikóva, **Agni Baka**, **Jonas Esser**, **Ceciel Pauls**, **Amrah Schotanus**, **Esther Maas**, **Ardine de Wit**, **Frank van Leth**, and **Matëj Kucera** I really enjoyed working and learning from you all!

My paranymphs, **Rafael Portela** and **Gisela Miyamoto** thank you so much for being there for me during my PhD journey. **Rafael**, you have been so kind to accept me the way I am, a busy mind involved with lots of research. You have encouraged me to get myself together at difficult times and also challenged me each and every day to put myself out there and stand up for myself. **Gisela**, we bonded in such a way that I cannot even put it into words. I enjoyed running cost-effectiveness analyses and debugging Stata code with you! Your care for me as a friend and as a colleague, which is very much appreciated!

I would like to express my deepest thanks to my family for the love and support they have given me during my life. Even though we are miles apart I believe that we grew closer. This thesis has for sure some parts of you in it! Ao meu amado pai, **José Renato** meu eterno agradecimento por ter me apresentado ao universo da pesquisa científica. À minha amada mãe **Jussara** agradeço imensamente por ter me ensinado com seu exemplo diário de perseverança e vitalidade. Aos meus amados irmãos **Tayná**, **João Batista**, sobrinhas **Petra** e **Heloísa** e cunhados **Vanessa** e **Luiz** o reconhecimento de todo o carinho, apoio incondicional e amizade! A big thank you also goes to my friends, **Airton Stein**, **Leila Coffy**, and **Maritza Bleil**, because you inspired me to think through things and go for it!

CHAPTER 14

PhD Portfolio

For the PhD thesis preparation, the courses below were followed as part of the Training Program of the Amsterdam Public Health Research Institute.

Courses	**Date achieved**	**EC**
Advanced Statistics (Vrije Universiteit Amsterdam)	Dec 2018	6.00
Scientific Project Management (Vrije Universiteit Amsterdam)	Oct 2019	0.50
Causal Inference and Propensity Score Methods (WR87 EpidM)	Jan 2021	3.00
Advanced Academic writing for PhD researchers (Vrije Universiteit Amsterdam)	Feb 2021	3.00
University Teaching Qualification (BKO, Vrije Universiteit Amsterdam)	Apr 2021	5.00
Career Days Vrije Universiteit Amsterdam	Nov 2021	0.50
Introduction to Probability and Data with R	Dec 2021	2.00
Research Integrity (VU Medical Center)	May 2022	2.00
BROK certification (NFU)	April 2022	1.00
Research related		
APH seminars, monthly department meetings, journal clubs, NVTAG meetings	–	2.00
ISPOR Barcelona 2018 (poster presentation)	Nov 2018	2.00
ISPOR Copenhagen 2019 (poster presentation)	Nov 2019	2.00
ERC EuroQol Meeting 2022 (paper presentation)	Apr 2022	1.00
HTAi Utrecht 2022 (poster presentation)	Jun 2022	2.00
Total number of ECTs		**32**

CHAPTER 15

List of publications

1. **Ben ÂJ**, van Dongen JM, Finch AP, Alili ME, Bosmans JE. To what extent does the use of crosswalks instead of EQ-5D value sets impact reimbursement decisions? A Simulation Study (ahead of print). *Eur J Health Econ*. Published online November 13, 2022. https://doi.org/10.1007/s10198-022-01539-6
2. **Ben ÂJ**, van Dongen JM, Alili ME, et al. The handling of missing data in trial-based economic evaluations: should data be multiply imputed prior to longitudinal linear mixed-model analyses? *Eur J Health Econ*. Published online September 26, 2022. doi:10.1007/s10198-022-01525-y
3. Souza TO de, **Ben ÂJ**, Dongen JM van, Bosmans JE, Cunha-Filho JSL da. Effectiveness and Cost-effectiveness of Minimal Ovarian Stimulation in-vitro Fertilization versus Conventional Ovarian Stimulation in Poor Responders: Economic Evaluation Alongside a Propensity Score Adjusted Prospective Observational Study. *JBRA Assist Reprod*. Published online September 15, 2022. doi:10.5935/1518-0557.20220025
4. Pellekooren S, **Ben ÂJ**, Bosmans JE, et al. Can EQ-5D-3L utility values of low back pain patients be validly predicted by the Oswestry Disability Index for use in cost-effectiveness analyses? *Qual Life Res*. Published online January 17, 2022. doi:10.1007/s11136-022-03082-6
5. Reinders I, Visser M, Jyväkorpi SK, **Ben ÂJ**, et al. The cost effectiveness of personalized dietary advice to increase protein intake in older adults with lower habitual protein intake: a randomized controlled trial. *Eur J Nutr*. 2022;61(1):505-520. doi:10.1007/s00394-021-02675-0
6. Miyamoto GC, **Ben ÂJ**, Bosmans JE, et al. Interpretation of trial-based economic evaluations of musculoskeletal physical therapy interventions. *Braz J Phys Ther*. 2021;25(5):514-529. doi:10.1016/j.bjpt.2021.06.011
7. Stegwee SI, **Ben ÂJ**, Alili ME, et al. Cost-effectiveness of single-layer versus double-layer uterine closure during caesarean section on postmenstrual spotting: economic evaluation alongside a randomised controlled trial. *BMJ Open*. 2021;11(7):e044340. doi:10.1136/bmjopen-2020-044340
8. van Dongen JM, **Ben ÂJ**, Finch AP, et al. Assessing the Impact of EQ-5D Country-specific Value Sets on Cost-utility Outcomes. *Medical Care*. 2021;59(1):82-90. doi:10.1097/MLR. 0000000000001417
9. Stein AT, **Ben ÂJ**, Pachito DV, Cazella SC, van Dongen JM, Bosmans JE. Digital Health Technology Implementation: Is It Effective in a Healthy Healthcare Perspective? In: Tevik Løvseth L, de Lange AH, eds. *Integrating the Organization of Health Services, Worker Wellbeing and Quality of Care: Towards Healthy Healthcare*. Springer International Publishing; 2020:197-220. doi:10.1007/978-3-030-59467-1_9
10. **Ben ÂJ**, Souza CF de, Locatelli F, et al. Health-related quality of life Associated with Diabetic Retinopathy in Patients of a Public Primary Care Service in Southern Brazil. forthcoming. *Archives of Endocrinology and Metabolism*. Published online 2019.
11. **Ben ÂJ**, Jelsma JGM, Renaud LR, et al. Cost-Effectiveness and Return-on-Investment of the Dynamic Work Intervention Compared With Usual Practice to Reduce Sedentary Behavior. *J Occup Environ Med*. 2020;62(8):e449-e456. doi:10.1097/JOM.0000000000001930
12. Boff TA, Pasinato F, **Ben ÂJ**, Bosmans JE, van Tulder M, Carregaro RL. Effectiveness of spinal manipulation and myofascial release compared with spinal manipulation alone on health-related outcomes in individuals with non-specific low back pain: randomized controlled trial. *Physiotherapy*. 2020;107:71-80. doi:10.1016/j.physio.2019.11.002

13. **Ben ÂJ**, Finch AP, Dongen JM van, et al. Comparing the EQ-5D-5L crosswalks and value sets for England, the Netherlands and Spain: Exploring their impact on cost-utility results. *Health Economics*. 2020;29(5):640-651. doi:10.1002/hec.4008

14. Stegwee SI, van der Voet LF, **Ben ÂJ**, et al. Effect of single- versus double-layer uterine closure during caesarean section on postmenstrual spotting (2Close): multicentre, double-blind, randomised controlled superiority trial. *BJOG*. 2021;128(5):866-878. doi:10.1111/1471-0528.16472

15. Dongen JM van, Alili ME, Varga AN, **Ben ÂJ**, et al. What do national pharmacoeconomic guidelines recommend regarding the statistical analysis of trial-based economic evaluations? *Expert Review of Pharmacoeconomics & Outcomes Research*. 2020;20(1):27-37. doi:10.1080/14737167.2020.1694410

16. **Ben ÂJ**, Neyeloff JL, de Souza CF, et al. Cost-utility Analysis of Opportunistic and Systematic Diabetic Retinopathy Screening Strategies from the Perspective of the Brazilian Public Healthcare System. *Appl Health Econ Health Policy*. 2020;18(1):57-68. doi:10.1007/s40258-019-00528-w

17. Lenzi H, **Ben ÂJ**, Stein AT. Development and validation of a patient no-show predictive model at a primary care setting in Southern Brazil. *PLOS ONE*. 2019;14(4):e0214869. doi:10.1371/journal.pone.0214869

18. Rosses APO, **Ben ÂJ**, Souza CF de, et al. Diagnostic performance of retinal digital photography for diabetic retinopathy screening in primary care. *Fam Pract*. 2017;34(5):546-551. doi:10.1093/fampra/cmx020

19. **Ben ÂJ**, Lopes JMC, Daudt CVG, Pinto MEB, Oliveira MMC de. Towards competency-based education: building the Family Medicine clerkship blueprint. *Rev Bras Med Fam Comunidade***.** 2017;12(39):1-16. doi:10.5712/rbmfc12(39)1354

20. Klein LF, Rigo SJ, Cazella SC, **Ben ÂJ**. *An Application for Mobile Devices Focused on Clinical Decision Support: Diabetes Mellitus Case. In: Lindgren H, Paz JFD, Novais P, et al., eds. Ambient Intelligence-Software and Applications – 7th International Symposium on Ambient Intelligence (ISAmI 2016)***.** Advances in Intelligent Systems and Computing. Springer International Publishing; 2016:57-65. doi:10.1007/978-3-319-40114-0_7

21. Cazella SC, Feyh R, **Ben ÂJ**. A Decision Support System for Medical Mobile Devices Based on Clinical Guidelines for Tuberculosis. In: Ramos C, Novais P, Nihan CE, Rodríguez JMC, eds. *Ambient Intelligence – Software and Applications*. Advances in Intelligent Systems and Computing. Springer International Publishing; 2014:217-224. doi:10.1007/978-3-319-07596-9_24

22. Souza CF de, **Ben ÂJ**, Schneider SMB, Nascimento BP, Neumann CR, Oliveira FJAQ de. The importance of programmatic health actions in tuberculosis control: experience of a Primary Health Care Service in Porto Alegre, Rio Grande do Sul, Brazil. *Clinical and Biomedical Research*. 2014;34(2). Accessed November 1, 2022. https://seer.ufrgs.br/index.php/hcpa/article/view/46878

23. **Ben ÂJ**, Neumann CR, Mengue SS. Teste de Morisky-Green e Brief Medication Questionnaire para avaliar adesão a medicamentos. *Rev Saúde Pública*. 2012;46:279-289. doi:10.1590/S0034-89102012005000013

CHAPTER 16

About the author

Ângela Jornada Ben was born on July 12th, 1980, in Passo Fundo, Southern Brazil. She graduated in Medicine in 2005 at the University of Passo Fundo. Shortly after her graduation, she worked as a medical doctor in a small Brazilian countryside town for a year. In 2007, she started medical residency in Family Medicine at the Hospital de Clínicas de Porto Alegre where she completed her training in 2011. Also in 2011, she earned her master's degree in Epidemiology from the Universidade Federal do Rio Grande do Sul (UFRGS) after studying the reliability and performance of questionnaires used to evaluate adherence to hypertensive treatment. In 2011, she started as a faculty at the Department of Collective Health of the Federal University of Health Sciences of Porto Alegre (UFCSPA) where she worked until 2016 supervising health sciences students and teaching epidemiology, family medicine, and public health. She also used to share her time working as a Family Doctor at the Brazilian Public Health System (SUS) and at the Telemedicine Government Program (TelessaúdeRS) supporting clinical decision-making by medical doctors working at SUS. Since August 2016, she is living with her partner in Amsterdam, the Netherlands. In the meantime, she worked on her PhD thesis, which she successfully defended in 2017 at the UFRGS Postgraduate Program in Epidemiology, Brazil. The focus of her thesis was to evaluate the cost-effectiveness of implementing Diabetic Retinopathy Screening Strategies in the SUS. To further develop her knowledge and skills in Health Economics and Health Technology Assessment, she went for a second PhD in the Department of Health Sciences at Vrije Universiteit Amsterdam. In 2022, she started working part-time as a research fellow at the Amsterdam Institute for Global Health & Development (AIGHD) to investigate the economic and health burden of Long COVID in Kenya. She is also working part-time as a postdoctoral researcher in the Department of Health Sciences at Vrije Universiteit Amsterdam to investigate the economic impact of Long COVID in The Netherlands.