Mapping clinical outcomes to generic preference-based outcome measures: development and comparison of methods

Mónica Hernández Alava,[1*] Allan Wailoo,[1] Stephen Pudney,[1] Laura Gray[1] and Andrea Manca[2]

[1]School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK
[2]Centre for Health Economics, University of York, York, UK

*Corresponding author monica.hernandez@sheffield.ac.uk

Declared competing interests of authors: none

Published June 2020
DOI: 10.3310/hta24340

Scientific summary

Developing and comparing mapping methods

Health Technology Assessment 2020; Vol. 24: No. 34
DOI: 10.3310/hta24340

NIHR Journals Library www.journalslibrary.nihr.ac.uk

Scientific summary

Background

Cost-effectiveness analysis using quality-adjusted life-years as the measure of health benefit is commonly used to aid decision-makers in health systems the world over. However, to calculate the quality-adjusted life-year benefits of different health technologies, policies or other types of interventions, it is necessary for clinical studies to include the correct types of outcome measures. Specifically, there are a range of preference-based measures that can be used. These preference-based measures comprise a survey instrument that is used to describe the health of an individual and a set of values for the health states that can be described. These values are typically derived from a sample of the general population using valuation methods that have some basis in economic theory. However, frequently, preference-based measures are entirely absent from clinical studies, or preference-based measures inappropriate for the setting are used, or the preference-based measure data from the studies are insufficient for the calculation of cost-effectiveness. In such circumstances, 'mapping' can bridge the evidence gap. Mapping entails estimating a statistical relationship, usually between clinical or other non-preference-based measures that have been used in clinical studies and the required preference-based measure for the economic analysis. To accomplish this, a different data set needs to be found in which both the required preference-based measure absent from the clinical study and the clinical or other non-preference-based measures used in the clinical study are present. Mapping is a widely employed method. However, mapping studies have often used methods that perform poorly, yield biased results and have a substantial impact on estimates of cost-effectiveness.

Objectives

- Further develop existing methods for mapping and develop new methods appropriate to the characteristics of health utility data.
- Develop methods for assessing mapping models for use in economic evaluation.
- Test the performance of mapping methods in data sets with differing characteristics, including a range of target preference-based measures.
- Develop methods for mapping between different preference-based measures that are specifically designed to allow the estimation of health benefits in one outcome measure or the other.
- Produce commands that allow other researchers to apply the statistical methods developed using standard statistical software.
- Identify key areas for further research.

Methods

Methods development

Health utility data are characterised by complex distributions that pose challenges for conventional statistical methods.

There are two approaches to mapping:

- direct mapping
    - one-step process
    - mapping specific to each country's utility
    - needs responses across the full range of disease severity.

- indirect mapping
    - two steps
    - can use the same model for different countries
    - needs enough responses at all levels in each dimension.

We developed and tested the performance of a series of mixture-model-based approaches for direct mapping in which the dependent variable is the health utility value. The adjusted limited dependent variable mixture model was originally developed in the context of the EuroQoL-5 Dimensions, three-level version, for the UK. It is based on mixtures of bespoke distributions that reflect the limits to the EuroQoL-5 Dimensions, three-level version, distribution at full health, at the worst health state and the substantial gap between full health and the next feasible health state. Other preference-based measures have less pronounced gaps than the EuroQoL-5 Dimensions, three-level version (using the UK value set), but they are all limited at both the top and the bottom of the health distribution. At the top, they are bounded by a common value, 1, representing full health. At the bottom, they are bounded at values that differ according to the worst health state defined by each preference-based measure. Whether or not there are large proportions of observations at any of these boundary points is also dependent on the specific characteristics of the patients of interest and the preference-based measure instrument. In this work, we further developed the adjusted limited dependent variable mixture model to allow options that reflect these different characteristics across preference-based measures: different boundaries at the bottom of the health utility distribution and the option to include/exclude the gap between full health and the next value in the health utility distribution using the gap relevant to each specific preference-based measure.

We developed beta-based mixture models that are a generalisation of the truncated inflated beta regression. The beta distribution is defined in the zero–one domain and does not allow observations at the boundaries. The general model developed here transforms the health utility data to the zero–one range so that the beta distribution can be used. It also allows observations at the boundaries via a choice of (1) adding a small amount of noise to the boundary observations if they are small in number or (2) allowing the inclusion of probability masses at those boundaries. Inclusion of the gap between full health and the next utility value, as well as a mass point at this value, are also options.

The developed direct approach models are also compared with standard indirect method approaches. Indirect methods, also known as response mapping, use a two-stage approach. First, the responses to the preference-based measure descriptive system are modelled. The expected utility value is calculated as a second step. These methods were tested where available data sets allowed them to be estimated.

We also developed an indirect approach for the case of mapping between preference-based measures. The mapping methods described above are designed to predict a preference-based measure from a set of non-preference-based measures used as explanatory variables. There is no necessity to map in the opposite direction. Therefore, mapping between preference-based measures requires further extension of methods to allow mapping in either direction in a mutually consistent way. The model was developed for mapping between the EuroQoL-5 Dimensions, three-level version, and EuroQoL-5 Dimensions, five-level version, but the general model can be used for any preference-based measures. We designed an approach that is as flexible as possible to avoid imposing unnecessary restrictions that could lead to inconsistent estimates. The structural parts of the model are allowed to differ between the preference-based measures, thus permitting post-estimation testing of hypotheses regarding the preference-based measures. The bivariate distribution between each pair of responses is specified using copulas, which allows the patterns of associations to differ across different dimensions of the preference-based measures and the strength of the association to differ at different parts of the health distribution. The model also relaxes the normality assumption underlying ordinal equations and incorporates a random latent factor to reflect individual-specific effects affecting the individual's responses across all dimensions of both preference-based measures.

There have been claims that mapping underestimates the degree of uncertainty for health state values. We demonstrate the difference between uncertainty and variability in the context of mapping models. We propose the use of graphical representations to compare like with like, that is, (1) the predictions from the models with the conditional sample means and (2) the distribution implied by the estimated model with the distribution of the sample data. We show how these can be used to judge model performance and to assess uncertainty for use in cost-effectiveness analyses.

In addition to developing mapping models, we also considered other methodological issues around mapping. First, we examined the extent of conflicts between the orderings of health states in a case study of the EuroQoL-5 Dimensions, three-level version, and EuroQoL-5 Dimensions, five-level version. Two instruments are monotonic if both order two health states unambiguously in the same way. Substantial failures of monotonicity present problems for mapping. Second, we investigated the issue of measurement error. There are many potential sources of measurement error in the health outcomes used in mapping: some are present in the responses to the variables being modelled in the mapping study, some are present when those same measures are used in clinical studies. Furthermore, we investigated the consequences of distributional mismatch between the trial target population and the population used for mapping as an additional source of bias.

Data

For mapping from non-preference-based measures to preference-based measures, seven case studies provided 15 data sets for estimation of mapping models. These studies were from patients with head injury, breast cancer (two case studies), asthma, heart disease, knee surgery and varicose veins. All four of the most widely used preference-based measures were considered in these case studies (though we concentrate on the two variants of the more commonly encountered EuroQoL-5 Dimensions): EuroQoL-5 Dimensions, three-level version ($n = 11$), EuroQoL-5 Dimensions, five-level version ($n = 2$), Short Form questionnaire-6 Dimensions ($n = 1$) and Health Utility Index Mark 3 ($n = 1$). Studies ranged in terms of sample size from 852 to 136,327 and were collected from randomised clinical trials, disease registries, bespoke patient survey studies and the UK NHS Patient Reported Outcome Measures (PROMs) programme.

For mapping between generic preference-based measures we used data from FORWARD, the National Databank for Rheumatic Diseases, which included both the EuroQoL-5 Dimensions, three-level version, and the EuroQoL-5 Dimensions, five-level version, in its 2011 wave ($n = 4856$).

Results

Figure a summarises the methods and variants tested in the case studies.

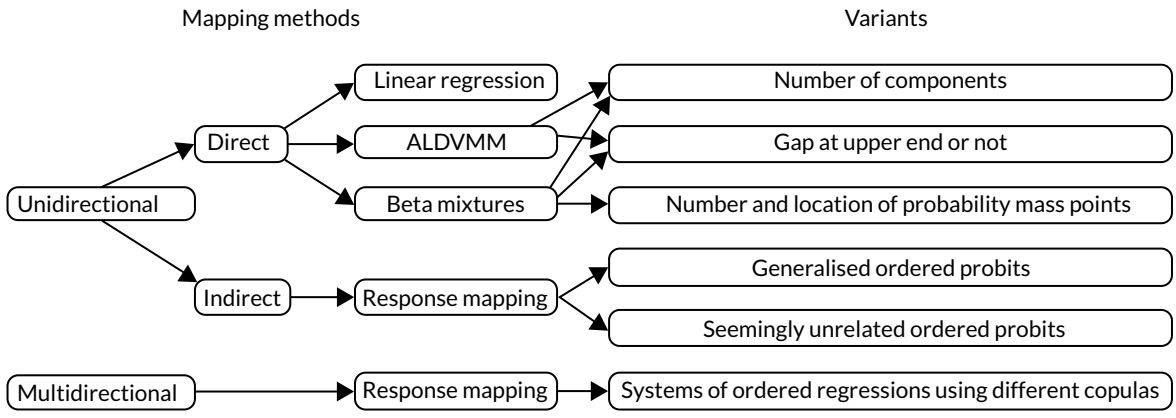

FIGURE a Summary of methods and variants tested. ALDVMM, adjusted limited dependent variable mixture model.

We demonstrate that there are problems with commonly used standard models. Linear regression is shown to be theoretically inappropriate for bounded data. Two-part models are designed to be applied to data which have a large proportion of observations at full health, but the approach is not flexible enough to deal with other challenges posed by health utility data. We show the importance of using appropriate model selection criteria to expose poor performance. In the case studies that examine methods for mapping from non-preference-based measures, we found that linear regression does not perform well, as predicted theoretically.

Flexible, direct mapping methods based on different variants of mixture models, with appropriately specified underlying distributions, perform very well for all preference-based measures, but the precise form is important. The case studies show that a minimum of three components are required. Covariates representing disease severity are required as predictors of component membership in all cases. Beta-based mixture models show similar performance to the adjusted limited dependent variable mixture model approaches but generally necessitate detailed consideration of the number and location of probability masses owing to the inability of the beta distribution to accommodate observations at the boundary values. Even for preference-based measures for which the gap between full health and the next feasible health state is much less pronounced than for the EuroQoL-5 Dimensions, three-level version (UK value set), explicitly allowing for this gap in the statistical model matters. Results were more variable regarding the optimal number of components and whether or not more than four could be estimated given the data, inclusion of probability masses at different points (upper and lower boundaries of the preference-based measure as well as the truncation point if a gap is included), covariates, their form and whether they appear within the components or in the component probabilities. Good practice needs to be followed in estimation to ensure that the models have genuinely converged.

Response mapping methods did not perform as well as direct mapping approaches in the three case studies in which these were applied. In some cases, response mapping could not be estimated because of lack of observations in each response category of the descriptive system of the preference-based measure.

We used the data and estimated models from six case study examples to demonstrate that mapping, from simple or complex direct methods or from indirect methods, does not underestimate uncertainty.

Econometric modelling based on a flexible mixture–copula specification response mapping model revealed significant differences between patients' responses to the three-level and five-level versions of the EuroQoL-5 Dimensions descriptive system. These differences were particularly striking for the mobility and pain domains, which are the most important dimensions of EuroQoL-5 Dimensions for patients with rheumatoid disease, the disease area of the estimation study. With regard to the health utility values, EuroQoL-5 Dimensions, five-level version, values had a tendency to be systematically higher than those of the EuroQoL-5 Dimensions, three-level version. We showed in an economic evaluation (of rheumatoid arthritis) case study that, as a direct consequence of these differences, the magnitude of the incremental cost-effectiveness ratio increased by > 100% in some cases when using the EuroQoL-5 Dimensions, five-level version, instead of the EuroQoL-5 Dimensions, three-level version, as used in the original economic evaluation.

We produced code in Stata® versions 14 and 15 (StataCorp LP, College Station, TX, USA) that allows the implementation of the methods developed as part of this project. Specifically, we wrote the Stata commands `aldvmm` and `betamix` to allow analysts to estimate mixture models appropriate for health utility data. We also wrote the Stata command `bicop`, which allows estimation of a simplified version of the flexible mixture-copula specification response mapping model. The command `eq5dmap` maps between the EuroQoL-5 Dimensions, three-level version, and EuroQoL-5 Dimensions, five-level version, in both directions from either the five-item health descriptions or the (exact or approximate) health utility score.

Conclusions

Mapping requires appropriate methods to yield reliable results. The appropriate methods should be judged by reference to (1) comparisons of the predictions with means of the data, grouped according to at least one of the conditioning variables, and (2) comparisons of the distribution of the data with the distribution of the data implied by the estimated model inter alia.

Widely used methods such as linear regression are not appropriate theoretically. Results from case studies clearly align with this position.

More flexible methods developed specifically for the purpose of unidirectional mapping show that close-fitting results can be achieved. The approaches based on mixture models that were developed here, namely the adjusted limited dependent variable mixture model and the beta-based mixture, are recommended for all preference-based measures. The precise form of these model types is important and were developed specifically to reflect the idiosyncrasies of health utility data. Some features are universally required but other features must be assessed on a case-by-case basis.

Case studies draw heavily on the EuroQoL-5 Dimensions because this is the most widely used preference-based instrument in UK policy and there are far fewer suitable data sets that use the Short Form questionnaire-6 Dimensions or Health Utility Index Mark 3.

Response mapping from non-preference-based measures to preference-based measures could not be undertaken for several case studies because of the lack of coverage, though this itself is revealing in relation to the feasibility of response mapping of preference-based measures with more complex descriptive systems.

The response mapping model developed specifically for multidirectional mapping performed very well. The two-step process of the response mapping approach was very useful in this case because it separates the responses to the descriptive system from the utility values attached to health states and, therefore, was able to uncover important differences between both versions of the EuroQoL-5 Dimensions.

We found that lack of monotonicity, distributional mismatch between the clinical study population and the population of the mapping study, and measurement error are all potential sources of bias. Illustrative case studies and calibration models showed that these biases might be substantial.

Recommendations for future research

Further research is needed to gain a better understanding of the problems posed by lack of monotonicity and the extent to which it affects different preference-based measures and non-preference-based measures. Furthermore, future research should examine the use of monotonicity measures for informing mapping studies.

Additional research is needed to understand the likely size of the biases due to distributional mismatch between the trial target population and the population used for mapping.

Given the different potential sources of measurement error in the outcomes used in mapping studies, future research is needed to generate evidence of how and when to adjust for measurement error in mapping.

Future research should concentrate on developing more flexible unidirectional response mapping models. Incorporating some of the features of the multidirectional mapping model developed here is one possible direction but others should also be explored.

Funding

This project was funded by the National Institute for Health Research (NIHR) Health Technology Assessment programme and will be published in full in *Health Technology Assessment*; Vol. 24, No. 34. See the NIHR Journals Library website for further project information. This project was also funded by a Medical Research Council grant (MR/L022575/1).

Health Technology Assessment

HTA/HTA TAR

ISSN 1366-5278 (Print)

ISSN 2046-4924 (Online)

Impact factor: 3.819

Health Technology Assessment is indexed in MEDLINE, CINAHL, EMBASE, the Cochrane Library and Clarivate Analytics Science Citation Index.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Technology Assessment* journal

Reports are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

Health Technology Assessment (HTA) research is undertaken where some evidence already exists to show that a technology can be effective and this needs to be compared to the current standard intervention to see which works best. Research can evaluate any intervention used in the treatment, prevention or diagnosis of disease, provided the study outcomes lead to findings that have the potential to be of direct benefit to NHS patients. Technologies in this context mean any method used to promote health; prevent and treat disease; and improve rehabilitation or long-term care. They are not confined to new drugs and include any intervention used in the treatment, prevention or diagnosis of disease.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

This report

The research reported in this issue of the journal was funded by the HTA programme as project number 15/141/09. The contractual start date was in September 2014. The draft report began editorial review in May 2019 and was accepted for publication in November 2019. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).