Mapping Patient-Reported Outcomes to EQ-5D Utilities in Hand Surgery: Development and Evaluation of Algorithms

Luke Geoghegan L[1], Conrad J Harrison[2], Shehan Hettiaratchy [3], Jeremy N Rodrigues[4,5]

[1] *Department of Surgery and Cancer, Imperial College London. London, UK*

[2] *Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford. Oxford, UK*

[3] *Department of Plastic Surgery, St Mary's Hospital, Imperial College Healthcare NHS Trust. London, UK.*

[4] *Department of Plastic and Reconstructive Surgery, Stoke Mandeville Hospital. Aylesbury, UK.*

[5] *Warwick Clinical Trials Unit, University of Warwick. Coventry, UK.*

Corresponding author
Prof. Jeremy Rodrigues
Warwick Clinical Trials Unit
University of Warwick
Coventry
UK
Tel: +44 (0) 2476 574 880
Email: jeremy.rodrigues1@nhs.net

Declaration of competing interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding statement: Conrad J. Harrison is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (NIHR300684). Jeremy N. Rodrigues is funded by a NIHR postdoctoral fellowship (PDF-2017-10-075). This document presents independent research funded by the NIHR. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Category: Original article

ABSTRACT

Health state utility values (HSUVs) describe the desirability of health states and are needed to calculate quality-adjusted life years. Traditionally, HSUVs are derived using generic measures such as the EQ-5D. No hand-specific measures can be used to derive HSUVs. This study aimed to develop and evaluate mapping algorithms to estimate the EQ-5D utility index from the Patient Evaluation Measure (PEM). Matched PEM and EQ-5D-5L data from 2043 patients were used to develop mapping algorithms. Transfer-to-utility models were developed and assessed through 1000-fold bootstrap validation. The Tobit censored regression model had the lowest mean absolute error of 0.127. We present a range of mapping algorithms to derive HSUVs from a commonly used hand-specific PROM. The population-level prediction error of our best performing models was clinically relevant. This suggests that mapping hand-specific PROMs onto generic preference-based measures may not be appropriate. Future work could seek to develop hand-specific preference-based measures.

Level of evidence: II

INTRODUCTION

Budget constrained healthcare systems must operate equitably to optimise outcomes for patients. The current tax-funded, free at the point of use healthcare model adopted by the UK National Health Service (NHS) requires complex resource allocation decisions to be made to ensure financial stability (Barr et al., 2014). Policymakers must ensure that appropriate valuation techniques are used to appraise interventions.

Quality-adjusted life years (QALYs) have emerged as a metric that enables life years to be adjusted to reflect health experienced during those life years (Lizan Tudela et al., 2009). The quality component of QALYs is known as the ‘utility value’ associated with a given health state. Utilities are typically scaled from 0 (equivalent to instant death) to 1 (equivalent to full health) (Raisch, 2000). The number of QALYs relating to a specific health state is expressed as the utility value of a given health state multiplied by the length of time in a given health state. Multi-attribute utility scales, such as the EuroQol five dimension (EQ-5D), are commonly used to estimate utilities. The EQ-5D has generic health classification system and a set of utility values assigned to each health state described (Herdman et al., 2011). It is widely accepted by policy makers and is the preferred preference-based instrument for technology appraisal in the UK (NICE, 2013).

Where EQ-5D utilities for a health state are not available, NICE recommend that utilities are derived by mapping available outcome measures onto preference-based PROMs such as the EQ-5D (NICE, 2013). The use of mapping algorithms in technology appraisal has grown substantially. A recent systematic review identified 144 mapping studies reporting 190 models from 110 distinct source outcome measures (Dakin et al., 2018). Mapping to predict EQ-5D utilities is of particular relevance in hand surgery, where hand-specific outcome measures predominate.

Less than 5% of published studies reported the use of preference-based measures (such as the EQ-5D), meaning cost-utility analyses cannot be performed using existing data (Lloyd-Hughes et al., 2019).

The paucity of primary utility data for hand conditions may mean that interventions are undervalued when cost-utility analyses are performed. This study aimed to develop and evaluate mapping models in the prediction of EQ-5D-5L utility scores from a commonly used hand-specific PROM, the Patient Evaluation Measure (PEM). The use of mapping algorithms may increase the availability of EQ-5D data available for cost utility analysis, meaning that novel and existing interventions for hand conditions can be accurately appraised.

METHODS

This study was conducted and is reported in line with the Mapping onto Preference-based measures reporting Standards (MAPS) statement (Petrou et al., 2015). This study used data from the UK Hand Registry (UKHR). The UKHR is a national voluntary registry for all patients undergoing elective intervention for hand and wrist conditions. Patients included in the UKHR are asked to complete both the EQ-5D-5L and the Patient Evaluation Measure (PEM) at baseline and 3, 6, 9 and 12-months following intervention. All patients gave informed consent for data collected to be used for research in anonymised fashion prior to inclusion in the UKHR.

Estimation sample

All consecutive adult patients who entered the registry between 2012-2020 were assessed for eligibility. Patients with complete, paired EQ-5D-5L and PEM response scores were eligible for inclusion. As this study aimed to develop cross-sectional mapping models, all paired EQ-5D-5L and PEM response scores were pooled, giving a total of 4,052 paired responses. We accounted for data-clustering associated with repeated observations by calculating robust standard error estimates for reported model coefficients.

Source and target measures

The PEM is a widely used hand-specific PROM (Lloyd-Hughes et al., 2019). The PEM was originally developed as a 10-item PROM with each item scored on a 7-point Likert scale (Macey et al., 1995). A further item examining the duration of pain was added in 2001 (Dias et al., 2001). The UKHR captured the original 10-item PEM up until 2017, after which it began to capture the updated 11-item PEM. The original 10 and updated 11-item PEM are scored from 10-70 and 10-77 respectively, with higher scores indicating greater disability. To maximise the use of available data, we included both response sets from the 10 and 11-item PEM in the UKHR. We reverse-scaled PEM responses to improve interpretability, meaning responses to both the 10 and 11-item PEM were rescaled from 0-100 with higher scores indicating better function.

The EQ-5D is a generic, preference-based measure of health with five domains: mobility, self-care, usual activities, pain and discomfort and anxiety and depression. The EQ-5D was originally developed on a three-point Likert scale (EQ-5D-3L) (Dolan, 1997). Recently, a five-level response PROM (EQ-5D-5L) was developed (Herdman et al., 2011). This comprises the same health domains however these are assessed on a five-point Likert scale. The health states described by the EQ-5D-5L have been valued by the UK population (Devlin et al., 2018). The UKHR captures the EQ-5D-5L; associated utilities were calculated with the UK value set using the R package ‘eq5d.’

Models

Transfer to utility (TTU) models were evaluated in the present study. These models, also known as direct utility mapping models, aim to predict EQ-5D utilities from total PEM response scores. The distribution of EQ-5D utilities was examined and regression techniques were selected accordingly, see Table 1, (Hernandez and Wailoo, 2015). The distribution of EQ-5D scores observed in the present study is in line with previously reported data across a range of health states (including those experienced by patients with conditions affecting the upper limb) and demonstrates notable key characteristics, including: a mass of observations at full health (utility score of 1), a gap between full and intermediate health states and a non-normal distribution (Hernandez Alava et al., 2018; Valsamis et al., 2023).

Given these distributional properties, we considered multiple modelling approaches. Censored regression (Tobit) was employed to address the upper bound of EQ-5D utilities at 1, ensuring that predicted values remained within a plausible range. However, given the clear gap between full and intermediate health states, a two-part model was also used. This approach incorporated an initial logistic regression step to classify patients with a utility score of 1 separately from those with lower scores, followed by a linear regression for patients with utilities below 1. This strategy accounted for the clustering of data at full health while improving prediction accuracy for those with intermediate health states. Additionally, an adjusted limited dependent variable mixture model was applied to account for the non-normal EQ-5D distribution, mass of obervations at 1 and gap between full and intermediate health states. The observed distribution of EQ-5D response data and subsequent model selection is in keeping with contemporary EQ-5D mapping studies (Dakin et al., 2018), see Table 1 for full model descriptions. We did not evaluate the effect of covariates such as age, gender, condition and intervention to improve parsimony.

Validation

An external validation dataset was not used in the present study due to lack of data availability. Instead, we randomly split our dataset into training (80% of total data) and testing (20% of total data) samples. Models were developed using the training dataset, model fit was then assessed using the testing dataset. We used bootstrap resampling to measure model stability through a 1000-fold repeated random split (Henderson, 2005). This means our dataset was randomly split 1000 times into training and testing sets; each time our models were developed and assessed on different datasets. We then calculated overall model performance across the 1000 datasets.

We assessed Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE) between observed and predicted EQ-5D utility scores as primary measures of model performance. We averaged measures of model performance across the 1000 testing datasets. All statistical analyses were undertaken using R (R Core Team, 2013).

RESULTS

The final dataset included a total of 4503 paired EQ-5D-5L and PEM responses from 2043 patients. There were missing data (defined as incomplete matched EQ-5D-5L and PEM responses) for 2611 patients. Of the included patients, 72% were diagnosed with base of thumb osteoarthritis (n= 1472), 2% were diagnosed with cubital tunnel syndrome (n = 35) and 26% were diagnosed with Dupuytren's disease (n= 536). The median age of included patients was 65 years (52 to 78 years); 36% of patients were male (n= 744) and 64% were female (n= 1299).

The distribution of EQ-5D-5L index and reverse scaled PEM scores are demonstrated in Figures 1 and 2, respectively. The distribution of EQ-5D-5L index values demonstrates a mass of observations at 1 (equivalent to perfect health) with a gap in utility values between full (utility score 1) and intermediate (utility score 0.95) health states in line with the UK EQ-5D-5L valuation tariff (Devlin et al., 2018). The distribution of reverse scaled PEM scores is approximately symmetric with a slight left skew (absolute value skewness -0.275). Both EQ-5D-5L utility and reverse-scaled PEM scores demonstrated clustering by condition, see Figures 3 and 4.

Mean and standard deviation for each performance measure across the testing data subsets is presented in Table 2. The tobit (censored regression) model had the lowest MAE (0.127) followed by the two-part model (0.128). The Adjusted Limited Dependent Variable Mixture Model (ALDVMM) had the highest MAE (0.132), whereas the Tobit had the highest RMSE (0.184), see Table 2.

DISCUSSION

This study presents a range of transfer-to-utility mapping algorithms to derive health state utility values from the PEM, a commonly used hand-specific PROM. Mapping algorithms were derived using routinely collected data from patients undergoing intervention for elective hand conditions as part of the UKHR.

Despite the use of a large national database in conjunction with sophisticated regression techniques, the population-level prediction error of our best performing models is clinically relevant. Our best performing model, the Tobit censored regression model, had a mean absolute error of 0.127. Consider a patient with an actual EQ-5D utility score of 0.5. An error of 0.127 could produce an HSUV prediction of 0.373 or 0.627. Contextually, this represents the difference between an individual with severe pain and severe depression and an individual with moderate pain and moderate mobility problems (Devlin et al., 2018). Comparative model performance was not consistent across different performance metrics (MAE and RMSE), with the Tobit model having the lowest MAE whilst the cubic polynomial model had the lowest RMSE. This is likely due to differences in residual size and differences in error penalisation between RMSE and MAE. The overall model prediction accuracy is in keeping with similar mapping studies. Valsamis *et al*., report comparable mapping algorithm prediction accuracy in their study which used the Oxford Shoulder Score as their source measure: MAE range 0.136 to 0.156 (Valsamis et al., 2023). Interestingly, Valsamis et al. report a higher MAE for complex, bespoke EQ-5D mapping models (namely the ALDVMM) as compared to simple linear regression models. This mirrors the results of our study where bespoke EQ-5D mapping algorithms had a higher degree of error compared to simple linear regression models. This may be due to the fact that mapping does not account for variables which effect overall quality of life that are not captured by the clinical measure of interest. Further, the EQ-5D may not be appropriate for all conditions and patient cohorts of interest. In such circumstances, the development of condition-specific preference-based measures should be considered.

In contrast to Valsamis *et al*., the present study solely focused on EQ-5D-5L derived utility values. This is because the UK Hand Registry only collects EQ-5D-5L data from respondents. In their 2019 position statement, NICE recommend that the EQ-5D-3L be used for reference case-analyses due to concerns regarding the quality and reliability of the data used in the valuation study ((NICE), 2019). Where primary data is collected using the EQ-5D-5L, NICE advocate the use of a mapping algorithm to derive EQ-5D-3L utility values. We opted to develop mapping algorithms using the 5L value set available in the UKHR as previous work has demonstrated comparable model performance using a primary EQ-5D-5L value set and EQ-5D-3L crosswalk values (Valsamis et al., 2023).

The present study demonstrates that the PEM can be mapped to the EQ-5D-5L using transfer-to-utility mapping. Response mapping aims to predict the responses to individual EQ-5D questions rather than the resultant HSUV. Using country specific value sets, HSUVs can then be determined from the mapped EQ-5D health state. Response mapping aims to address the inherent limitations of transfer-to-utility mapping, however estimating response mapping models is computationally intensive and requires access to complete EQ-5D response data (Hernández Alava et al., 2020). Future mapping studies should investigate whether the PEM can be accurately mapped to EQ-5D-3L index values using response mapping. A range of hand-specific PROMs have been developed. Future work should also ascertain the accuracy of mapping algorithms developed using a range of hand-specific PROMs.

The clinically relevant error margin demonstrated in the present study may suggest that mapping the PEM onto generic preference-based measures may not be appropriate. The assumption that the EQ-5D is an appropriate measure for use in patients with hand conditions is questionable. Previous work has shown that the EQ-5D is not responsive to clinical deterioration in patients with rheumatoid arthritis (Payakachat et al., 2015). The poor responsiveness of generic preference-based measures such as the EQ-5D may mean that interventions for hand conditions are undervalued in cost-effectiveness analyses. This risks interventions being labelled ‘of limited clinical value’ when compared to treatments for other conditions such as knee arthritis. One solution to the poor responsiveness of generic preference-based measures is the use of condition-specific preference-based measures (CS-PBMs). CS-PBMs can be derived from existing PROMs which pose more relevant questions to the target population using established methods (Brazier et al., 2012). Future work should aim to develop a value set for an established hand-specific PROM. This would enable utility values to be determined for health states described by the hand-specific PROM and overcome the limitations of generic preference-based measures in hand surgery.

The present study is not without limitations. We did not have a separate external validation dataset and instead chose to perform bootstrap validation with 1000-fold test-training samples and validating models across these splits. Given that error is likely to increase in an external validation sample, this limitation does not detract from the principle study findings. We did not examine the effect of additional variables such as age and gender to improve the generalisability of models to datasets where such data are not available.

Our findings are applicable across surgical sub-specialties. We have demonstrated that responses to domain-specific PROMs can be mapped to EQ-5D utilities, albeit with a clinically relevant margin of error. The principal advantage of mapping is increased availability of data for cost-utility analyses in the absence of primary EQ-

5D data. However, analysts must interpret the results of mapped utility values with caution given the additional uncertainty associated with having to derive utility scores using responses to distinct outcome assessment tools. Analysts must consider whether the EQ-5D is an appropriate measure in the cohort of interest. This is of particular relevance in hand surgery where the impact of hand conditions may not impair health domains measured by the EQ-5D such as mobility. However, general surgical pathologies (such as colorectal cancer) may be associated with greater general disability and subsequent impairments in quality of life may be better represented in EQ-5D scores, meaning mapping may be appropriate. Where this is not the case, NICE advocate the use of condition-specific preference-based measures for technology appraisal (NICE, 2013). Efforts should therefore be concentrated on the development of condition-specific preference-based measures in relevant cohorts, rather than the development of mapping algorithms.

In conclusion, the mapping models developed and evaluated in this study enable decision analysts to obtain EQ-5D-5L HSUVs from the PEM where primary EQ-5D-5L data are not available. The error margin of our best performing model is clinically significant. Future work should aim to develop a value set for hand-specific PROMs.

FIGURE LEGENDS

Figure 1: EQ-5D-5L utility distribution
Figure 2: Reverse scaled PEM distribution
Figure 3: Density plot of EQ-5D-5L utilities stratified by condition
Figure 4: Density plot of reverse scaled PEM scores stratified by condition

REFERENCES

(NICE) NIfHaCE. Position statement on use of the eq-5d-5l value set for england (updated october 2019). 2019. https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l (26/05/2024).

Barr B, Bambra C, Whitehead M. The impact of nhs resource allocation policy on health inequalities in england 2001-11: Longitudinal ecological study. Bmj. 2014, 348: g3231.

Brazier JE, Rowen D, Mavranezouli I et al. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technol Assess. 2012, 16: 1-114.

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 Qual Life Outcomes. 2018, 16: 31.

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.

Dias JJ, Bhowal B, Wildin CJ, Thompson JR. Assessing the outcome of disorders of the hand. Is the patient evaluation measure reliable, valid, responsive and without bias? J Bone Joint Surg Br. 2001, 83: 235-40.

Dolan P. Modeling valuations for euroqol health states. Med Care. 1997, 35: 1095-108.

Henderson AR. The bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. Clin Chim Acta. 2005, 359: 1-26.

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.

Hernandez Alava M, Wailoo A, Grimm S et al. Eq-5d-5l versus eq-5d-3l: The impact on cost effectiveness in the united kingdom. Value Health. 2018, 21: 49-56.

Hernández Alava M, Wailoo A, Pudney S, Gray L, Manca A. Mapping clinical outcomes to generic preference-based outcome measures: Development and comparison of methods. Health Technol Assess. 2020, 24: 1-68.

Hernandez M, Wailoo A. Fitting adjusted limited dependent variable mixture models to eq-5d. Stata Journal. 2015, 15: 737-50.

Lizan Tudela LV, Paz S, Rodriguez JM, Gonzalez P. Quality adjusted life years (qalys) in economic evaluations of health technologies in spain: A review of the 2003-2009 literature. Value in Health. 2009, 12.

Lloyd-Hughes H, Geoghegan L, Rodrigues J et al. Systematic review of the use of patient reported outcome measures in studies of electively-managed hand conditions. J Hand Surg Asian Pac Vol. 2019, 24: 329-41.

Macey AC, Burke FD, Abbott K et al. Outcomes of hand surgery. Journal of Hand Surgery. 1995, 20: 841-55.

NICE. Nice process and methods guides. In: *Guide to the methods of technology appraisal 2013*, London, National Institute for Health and Care Excellence (NICE)

Copyright © 2013 National Institute for Health and Clinical Excellence, unless otherwise stated. All rights reserved., 2013.

Payakachat N, Ali MM, Tilford JM. Can the eq-5d detect meaningful change? A systematic review. Pharmacoeconomics. 2015, 33: 1137-54.

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.

R Core Team R. R: A language and environment for statistical computing. 2013.

Raisch DW. Understanding quality-adjusted life years and their application to pharmacoeconomic research. Ann Pharmacother. 2000, 34: 906-14.

Valsamis EM, Beard D, Carr A et al. Mapping the oxford shoulder score onto the eq-5d utility index. Qual Life Res. 2023, 32: 507-18.