Received: 4 December 2019 | Revised: 7 December 2021 | Accepted: 9 January 2022

DOI: 10.1002/hec.4503

RESEARCH ARTICLE

Health Economics WILEY

# An examination of machine learning to map non-preference based patient reported outcome measures to health state utility values

**Mona Aghdaee**[1] | **Bonny Parkinson**[1] | **Kompal Sinha**[2] | **Yuanyuan Gu**[1] | **Rajan Sharma**[1] | **Emma Olin**[1] | **Henry Cutler**[1]

[1]Macquarie University Centre for the Health Economy, Macquarie University, Sydney, New South Wales, Australia

[2]Department of Economics, Macquarie Business School, Macquarie University, Sydney, New South Wales, Australia

**Correspondence**
Mona Aghdaee, Macquarie University Centre for the Health Economy, Macquarie University, Level 1, 3 Innovation Road, Sydney, NSW 2109, Australia.
Email: mona.aghdaee@mq.edu.au

**Funding information**
Central Coast Local Health District (CCLHD), New South Wales, Australia
Open access publishing facilitated by Macquarie University, as part of the Wiley - Macquarie University agreement via the Council of Australian University Librarians.

**Abstract**

Non-preference-based patient-reported outcome measures (PROMs) are popular in health outcomes research. These measures, however, cannot be used to estimate health state utilities, limiting their usefulness for economic evaluations. Mapping PROMs to a multi-attribute utility instrument is one solution. While mapping is commonly conducted using econometric techniques, failing to specify the complex interactions between variables may lead to inaccurate prediction of utilities, resulting in inaccurate estimates of cost-effectiveness and suboptimal funding decisions. These issues can be addressed using machine learning. This paper evaluates the use of machine learning as a mapping tool. We adopt a comprehensive approach to compare six machine learning techniques with eight econometric techniques to map the Patient-Reported Outcomes Measurement Information System Global Health 10 (PROMIS-GH10) to the EuroQol five dimensions (EQ-5D-5L). Using data collected from 2015 Australians, we find the least absolute shrinkage and selection operator (LASSO) model out-performed all machine learning techniques and the adjusted limited dependent variable mixture model (ALDVMM) out-performed all econometric techniques, with the LASSO performing better than ALDVMM. The variable selection feature of LASSO was then used to enhance the performance of the ALDVMM in a hybrid model. Our analysis identifies the potential benefits and challenges of using machine learning techniques for mapping and offers important insights for future research.

**KEYWORDS**
econometrics, EQ-5D, machine learning, mapping, PROMIS, utility

## 1 | INTRODUCTION

Patient-reported outcome measures (PROMs) are being used more often in healthcare systems as funders increasingly seek value-based care. Non-preference based PROMs are increasingly included in clinical studies, health service management and

research. However, these measures cannot be used to estimate health state utility values (henceforth "utilities"), limiting their usefulness for economic evaluations. Mapping non-preference based PROMs to a multi-attribute utility instrument (MAUI), which can be used to estimate utilities, is one solution to this problem (Kearns et al., 2013).

Mapping is a statistical technique used to link outcomes from non-preference-based PROMs ("explanatory variables") to a MAUI using an alternative data source. The benefits of mapping have been acknowledged in a review of the UK's National Institute for Health and Care Excellence (NICE) appraisals conducted over 2004–2008 (Tosh et al., 2011), which found an increase in the use of utility mapping approaches, accounting for over a quarter of total submissions. Consequently, the updated NICE guidelines for 2013 recommended using mapping to estimate utilities in the absence of direct utility measures (National Institute for Health and Care Excellence, 2013). Mapping is also accepted by the Pharmaceutical Benefits Advisory Committee and the Medical Services Advisory Committee (MSAC) in Australia as an alternative approach for estimating utilities in economic evaluations (Department of Health, 2016; Medical Services Advisory Committee, 2016).

Existing literature on mapping health outcomes has adopted direct and indirect mapping approaches. The direct approach estimates utilities directly from explanatory variables. The indirect approach, also known as response mapping, first predicts the probabilities for each response to each MAUI question, and then uses relevant tariffs to convert them into utilities (Hernandez-Alava et al., 2014). The resulting algorithm can be applied to PROM data to estimate the associated utilities, and thus quality adjusted life years (QALYs) (Wailoo et al., 2017).

To investigate the validity of different mapping techniques, Brazier and colleagues undertook a systematic review of studies mapping between non-preference based PROMs, generic preference based measures and MAUIs (Brazier et al., 2010). They found most studies using the direct approach typically adopted the linear, ordinary least square (OLS) regression technique to predict health state utilities. This could result in inaccurate prediction of utilities since utilities are bounded at one and have a distribution skewed to the left (Ara & Brazier, 2008; Brazier et al., 2010; Crott & Briggs, 2010; Rowen et al., 2009). Other popular econometric techniques used to directly predict utilities include Tobit (Sullivan & Ghushchyan, 2006), generalizsed linear model (GLM) (Sharma et al., 2019), censored least absolute deviation (CLAD) (Kaambwa et al., 2006; Sullivan & Ghushchyan, 2006), and median regression (Wu et al., 2007). Each of these techniques are better suited to predict utilities than the OLS technique particularly in accommodating unique characteristics of utilities being bounded, and clustering at one. Specifically, the standard Tobit technique accounts for the bounded utilities but does not allow for a gap below the mass of observations at one found in preference-based measures (Sullivan & Ghushchyan, 2006). Certain families of GLM are able to accommodate flexible non-linear relationships but may produce inconsistent estimates when the link function is misspecified (Dakin et al., 2010). Median regression is more robust to outliers (Shaw et al., 2010) but does not consider utilities being bounded. The CLAD extends the median regression with the dependent variable constrained on a fixed interval (Powell, 1984). However, since cost-effectiveness analyses (the main reason for needing mapping exercises) are based on mean values, techniques based on medians are less useful. For indirect mapping, multinomial logit (MLOGIT), ordered logit (OLOGIT), and generalized ordered logit (GLOGIT) have been applied in the literature (Gray et al., 2006).

Recently, mixture models such as the mixture beta regression model (Betamix) and the adjusted limited dependent variable mixture model (ALDVMM) have been adopted in mapping studies as preferred techniques due to their flexibility and ability to accommodate multimodality (Basu & Manca, 2012; Gray & Hernandez-Alava, 2018; Hernandez-Alava & Wailoo, 2015; Hernandez-Alava et al., 2013; Khan & Morris, 2014; Yang, Wong, et al., 2019; Young et al., 2015). The Betamix is a two-part model (consisting of a multinomial logit and a beta mixture model), which allows estimation of multimodal dependent variables bounded in an interval (Gray & Hernandez-Alava, 2018) and has been shown to out-perform linear regression (Khan & Morris, 2014; Yang, Wong, et al., 2019). ALDVMM is a mixture model of adjusted Tobit-like distributions (Hernandez-Alava & Wailoo, 2015), which deals with utility data's distributional features and accounts for the multimodality. ALDVMM assumes that utilities can be modeled as a mixture of multiple components, each representing a cluster of respondents with similar utility scores. It combines multiple component distributions with a multinomial logit model of the probabilities of component membership. ALDVMM has been shown to perform better than other traditional econometric techniques used in the mapping literature (Gray & Hernandez-Alava, 2018).

While mapping has become a common practice in estimating utilities, the characteristics of health utilities may limit the accuracy of mapping algorithms. In addition to being bounded highly skewed (full health) (Brazier et al., 2010), utilities often have conditional distributions that are not easily accommodated by standard parametric distributions. For economic evaluations, it is imperative for these utility predictions to be accurate. The relationships between PROMs and MAUIs are commonly non-linear and involve complex interactions among explanatory variables. In standard econometric techniques used for mapping, selection of the distribution function and explanatory variables are based on prior knowledge of the clinical relationships between the variables for standard statistical tests. Moreover, the probabilistic distribution of the error terms is often not explicit and the relevant explanatory variables and their relationship with utilities are not immediately apparent. Failing to

specify these relationships appropriately will reduce the accuracy of the mapping algorithm. One way to avoid this potential problem is to use machine learning techniques for mapping.

The use of machine learning has increased in recent years in all areas of research (Athey & Imbens, 2019), including health economics. Applications include estimating the treatment effects of medical interventions (Kreif et al., 2015), analysis of prescribing patterns (Schilling et al., 2016), identifying thresholds and hierarchies in funding decisions (Schilling et al., 2017), and predicting healthcare costs (Konig et al., 2013). The key strengths of machine learning techniques compared to standard econometric techniques are prediction accuracy and parsimony as there is less requirement to impose parameters. Machine learning does not require prespecifying the probabilistic distribution of the error term, selecting explanatory variables, or assuming their inter-relationships that is, additive or multiplicative interactions of their effect on the conditional mean of the outcome as well as their linear or non-linear associations with the dependent variable (Varian, 2014). This is particularly useful when explanatory variables are numerous, and their significance and potential interactions are unknown. While it is not feasible to test all the possible combinations of explanatory variables with standard econometric techniques, machine learning has the advantage of using data driven techniques to determine the relationships between explanatory variables and outcome (Breiman et al., 1984; Strobl et al., 2009).

The objective of this study was to evaluate the performance of machine learning techniques for mapping non-preference based PROMs to MAUIs compared to standard econometric techniques. In the absence of preference based measures, mapping predicts utilities from non-preference based PROMs. Given the predicted utilities are used in economic evaluation and ultimately in funding decisions, producing a robust and appropriate mapping algorithm is crucial as the accuracy of a mapping technique affects the predicted utilities, and thus the estimated cost-effectiveness of an intervention. Thus, it is essential to compare the performance of commonly used econometric techniques with a selection of machine learning techniques, and choose the most accurate one (Yang, Devlin, & Luo, 2019).

One of the most popular and well-established PROM is the Patient Reported Outcomes Measurement Information System (PROMIS), developed by the National Institutes of Health in the United States in 2004. One of its three instrument types is the PROMIS short form Global Health 10 (PROMIS-GH10), which is a generic measure of health focusing on physical, mental and social well-being from the patient perspective (Cella et al., 2010; Hays et al., 2009). The PROMIS-GH10 is widely used across the world as the gold standard for patient-centered assessment. In Australia, New South Wales (NSW) Health has adopted PROMIS-GH10 as a key evaluation component of the NSW Health Integrated Care Strategy (Thompson et al., 2016). In the UK, the National Institute for Health Research has supported validating and calibrating PROMIS-GH10 for administration in clinical practices and research, in an attempt to unify the PROMs and shift toward a more patient-centered health system (Evans et al., 2018). Internationally, PROMIS-GH10 has been recommended as a core outcome measure in several clinical areas by the International Consortium for Health Outcomes Measurement (Nijagal et al., 2018; Salinas et al., 2016).

The growing preference toward patient reported outcomes has resulted in a rapidly expanding literature using PROMIS-GH10 to collect patient reported data. Since a commonly used measure in economic evaluations is the EuroQol five dimensions (EQ-5D-5L), this paper predicted utilities from the PROMIS-GH10 response using EQ-5D-5L as the target measure of mapping. The relationship between PROMIS-GH10 and EQ-5D-5L questions is not obvious and given the complexity of the possible interactions among the questions and different levels, there is potential to explore the latest techniques such as machine learning to improve mapping accuracy. This paper makes three important contributions to the literature. First, based on the techniques used in the existing literature, we used a range of econometric techniques including linear regression, Tobit, median regression, GLM, CLAD, Betamix, ALDVMM, and GLOGIT and machine learning techniques including classification and regression trees analysis (CART), bagged CART, random forests, Neural Networks (NN), quantile regression neural networks (QRNN), and least absolute shrinkage and selection operator (LASSO) to map from PROMIS-GH10 to EQ-5D-5L. To the best of our knowledge, this is the first study to apply multiple machine learning techniques to map non-preference based PROMs to a MAUI and compare them to econometric techniques. The only other study comparing the performance of econometric techniques to machine learning techniques was Park and Basu (2018), which assessed the predictive accuracy of these techniques in the context of risk-adjustment in the health insurance market. Second, capitalizing on our approach of comparing techniques, we combine the best performing machine learning technique (LASSO) and best performing econometric technique (ALDVMM) to propose a hybrid model for prediction. This enabled us to highlight the advantage of combining machine learning and econometric techniques for better outcomes particularly since LASSO as a prediction technique cannot produce a mapping algorithm. Finally, while most existing studies focused on mapping PROMIS-GH10 to EQ-5D-3L (Revicki et al., 2009; Thompson et al., 2017), we undertook the first mapping exercise to map from PROMIS-GH10 to EQ-5D-5L, which has greater sensitivity and covers a wider range of health states. We provide a mapping algorithm to predict EQ-5D-5L utilities when only PROMIS-GH10 data is collected but a health economic evaluation is desired.

The rest of this paper is organized as follows. The next section describes our data and the measures of performance followed by Section 3, where we discuss the methods. Section 4 presents the results and Section 5 concludes with a discussion.

## 2 | DATA

An online survey was conducted in February 2018 to collect responses to the PROMIS-GH10 and EQ-5D-5L instruments from a representative general population of 2015 Australians (Hays et al., 2009; Herdman et al., 2011). The PROMIS-GH10 consists of 10 questions about physical function, pain, fatigue, emotional distress, social health, and general perceptions of health. Each question measures the severity level ranging between one and five, except for pain which ranges from 0 to 10. Two summary scores of physical and mental health are derived from PROMIS-GH10 (Hays et al., 2009).

The five-level version of the EQ-5D has recently been introduced by the EuroQol Group to attain greater sensitivity to health states changes and a broader range of utilities than the previous three-level version (EQ-5D-3L) (Janssen et al., 2008). The EQ-5D-5L consists of five questions about mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has five levels from having no problems to having extreme problems (Herdman et al., 2011). The EQ-5D-5L utilities were estimated using Australian tariffs (Norman et al., 2017), (see tab. 1, Approach 5 of their paper). Demographic information on age, sex, state, postcode, and an optional response to the Charlson Comorbidity Index (CCI) was also collected (Chaudhry et al., 2005).

## 3 | METHODS

All statistical analyses were conducted in STATA 16 and R (Version 4.0.3). The mapping techniques used in this paper comply with the ISPOR Good Practices for Outcomes Research Task Force Report (Wailoo et al., 2017), and the Mapping onto Preference-Based Measures Reporting Standards (MAPS) checklist (Dakin et al., 2018; Petrou et al., 2015) (see Appendix A for details).

### 3.1 | Overview

We developed algorithms to predict the conditional mean of the target measure (here EQ-5D-5L utilities) from the observations of the source measure (here PROMIS-GH10). The predictions were then compared with the actual target measure observations to assess the accuracy of the algorithms.

In direct mapping, source measure or explanatory variables (here the PROMIS-GH10 items or summary scores) were directly mapped onto the target measure or dependent variable (here EQ-5D-5L utility values). In comparison, indirect mapping was performed in two stages: the responses to each dimension of the target measure (EQ-5D-5L dimensions: mobility, self-care, usual activity, pain and discomfort and depression and anxiety) were considered as the dependent variable; and then the predicted responses were combined using a relevant tariff to estimate utilities.

### 3.2 | Measures of model performance

The performance of prediction models was measured by in-sample cross-validation using a k-fold technique (Fushiki, 2011) for 10 folds. The dataset was randomly divided into $k = 10$ subsamples, of which $k\text{-}1 = 9$ subsamples were used as the estimation sample, and one subsample was used as the validation sample for testing the accuracy of the predictions. This process was repeated 10 times with each of the 10 subsamples used once as the validation data. The 10-fold cross-validation was performed for both machine learning and econometrics techniques to enable comparability.[1]

The predictive accuracy was determined by the degree to which the predicted utilities reflected the observed utilities. The primary measure of predictive accuracy was average Mean Absolute Error (MAE) after truncation across the validation subsamples (Wailoo et al., 2017). While MAE was used as the primary measure of the predictive accuracy of each technique, other measures of predictive accuracy were also reported, including the MAE before truncation, the Mean Squared Error (MSE) before and after truncation, the predicted mean utility, the predicted minimum utility, and the predicted maximum utility.[2]

The predicted mean utility was reported as this is often used in cost-effectiveness analyses, while the minimum and maximum utilities were reported to assess how the techniques performed in the extremes. Plots comparing the distribution of the observed versus predicted utilities were also presented to examine how each technique fits different parts of the distribution. It is important to note that the goodness of fit criteria was based on overall utility and does not reveal the prediction accuracy of the techniques relating to the underlying items. This does not affect the analysis as the objective was to assess the prediction accuracy relating to the overall utility for each respondent, which can be used in cost-effectiveness analyses. Due to the lack of data on the five-level version of EQ-5D-5L, no external dataset was available, thus, only internal cross-validation was applied in this study.

## 3.3 | Econometric techniques

### 3.3.1 | Direct mapping

In direct mapping, where the explanatory variables (here PROMIS-GH10 items or summary scores) were directly mapped onto the EQ-5D-5L utility values, seven econometric techniques were used. This included linear regression, Tobit, median regression, GLM, CLAD, Betamix, and ALDVMM.

The dependent variable (target measure) in estimating the linear regression, Tobit, median regression, GLM and CLAD techniques was disutilities (=1-utilities) and predictions were deducted from one to estimate utilities. We used the utilities as the dependent variable to estimate the Betamix and ALDVMM models.

Four models based on the sets of explanatory variables were specified as follows:

Set 1:

$$EQ\text{-}5D\text{-}5L_i = Max\left[f\left(\beta_0 + \beta_1 * PROMIS\text{-}GH10\,Physical\text{-}score_i + \beta_2 * PROMIS\text{-}GH10\,mental\text{-}score_i + \gamma_1 * Sex_i + \gamma_2 * Age_i + \gamma_3 * Age_i^2\right), 1\right] \quad (1)$$

Set 2:

$$EQ\text{-}5D\text{-}5L_i = Max\left[f\left(\beta_0 + \sum_{k=1}^{K} \delta_{ik} * PROMIS\text{-}GH10\_items\_cont_{ik} + \gamma_1 * Sex_i + \gamma_2 * Age_i + \gamma_3 * Age_i^2\right), 1\right] \quad (2)$$

Set 3:

$$EQ\text{-}5D\text{-}5L_i = Max\left[f\left(\beta_0 + \sum_{j=1}^{J} \delta_{ij} * PROMIS\text{-}GH10\_items\_cat_{ij} + \gamma_1 * Sex_i + \gamma_2 * Age_i + \gamma_3 * Age_i^2\right), 1\right] \quad (3)$$

Set 4:

$$EQ\text{-}5D\text{-}5L_i = Max\left[f\left(\beta_0 + \sum_{j=1}^{J} \delta_{ij} * PROMIS\text{-}GH10\_items\_cat_{ij} + \gamma_1 * Sex_i + \partial_{ij} * Age\_cat_{ij}\right), 1\right] \quad (4)$$

where $EQ\text{-}5D\text{-}5L_i$ represents the predicted utility for the individual $i$.

In set 1, EQ-5D-5L utilities were predicted using the physical (*PROMIS-GH10 physical-score*) and mental (*PROMIS-GH10 mental-score*) health summary scores of PROMIS-GH10 (as continuous variables), age, age squared, and sex. In set 2, all the PROMIS-GH10 questions as continuous variables, age, age squared, and sex were included. The set 3 of explanatory variables consisted of PROMIS-GH10 questions as categorical variables (*PROMIS-GH10_items_cat*), age, age squared, and sex; and for set 4, PROMIS-GH10 questions, age (*Age_cat*), and sex (all as categorical variables) were considered. The age categories were defined based on Australian Bureau of Statistics (ABS) age categories (Australian Bureau of Statistics, 2017). Sets 1 and 2 were selected according to Revicki et al. (2009). Sets 3 and 4 directly included PROMIS-GH10 items to take into account the ordinal nature of PROMIS-GH10 responses (Revicki et al., 2009).

In the estimation of GLM, the Modified Parks Test identified the family distribution of Poisson and log link for the EQ-5D-5L utilities (Manning & Mullahy, 2001). Results were reported with and without predicted utilities being truncated at one.

### 3.3.2 | Indirect mapping

In indirect mapping, the responses to each EQ-5D-5L question were the dependent variables, then the predicted responses were combined to predict utilities. As each question was modeled separately, each mapping algorithm consisted of five separate models. One set of explanatory variables was considered in indirect mapping, including PROMIS-GH10 questions as categorical variables, age, and sex.

In indirect mapping, as the dependent variables are categorical variables with discrete outcomes, one option would be the use of the ordered logit model (OLOGIT) to predict the probability of each response level. The OLOGIT has the advantage of accounting for the order of categorical responses to EQ-5D-5L questions. However, the OLOGIT relies on an assumption of proportional odds or parallel lines/slopes. It generates a set of binary response models for the different ordered categories, in which the intercepts are different, but the coefficients for the explanatory variables are the same. This leads to the cumulative probability curves for the different ordered categories having parallel slopes. If this assumption is violated, OLOGIT provides biased estimates. An alternative to OLOGIT is the multinomial logit model (MLOGIT). However, it does not consider the ordinal structure of the dependent variables.

In this paper, the generalized logit model (GLOGIT) was chosen over MLOGIT or OLOGIT as while it considers the ordinal structure of the dependent variable; it is less restrictive in relaxing the parallel slopes assumption (Long & Freese, 2006). The conditional probability of an observation belonging to class $m$, (for $m = 2$–$5$) can be written as:

$$\Pr\left(EQ\text{-}5D\text{-}5L\text{-}dimentions_i = m|X_i\right) = \frac{exp(X_i\beta_m)}{1 + \sum_{j=2}^{J} exp(X_i\beta_j)} \tag{5}$$

where $m$ denotes one of the five dimensions of EQ-5D-5L and $exp(X_i\beta_m)$ is normalized as 1 for the reference category.

GLOGIT generates several equations, each of them being a binary logistic regression that compares that group with a reference group, and each of them yields a probability that the observation falls into that category. Once these were obtained, individuals were assigned to one of the five levels using a Monte Carlo simulation approach where the predicted probabilities were compared to a random number from a uniform distribution. We ran 100 Monte Carlo simulations across the full sample. This approach is known to produce a more accurate distribution of responses in each dimension of EQ-5D-5L (Gray et al., 2006). Then the predicted responses were combined and utilities were calculated using the Australian EQ-5D-5L tariff (Gray et al., 2006; Long & Freese, 2006; Norman et al., 2017).

## 3.4 | Machine learning techniques

Supervised machine learning techniques are primarily concerned with building predictive models that performs well in predicting outcomes for yet unseen data. An important feature of these techniques making them suitable for mapping is their ability to incorporate a large set of variables in a non-linear pattern to improve prediction accuracy. We explored six supervised machine learning techniques to map from the PROMIS-GH10 to the EQ-5D-5L, including CART, bagging, random forests, NN, QRNN and LASSO. The choice of techniques was based on the relative advantage of each technique. For all the machine learning techniques, except for LASSO, the explanatory variables were not prespecified. Instead, the explanatory variables included was decided by the machine learning technique from the set of all the potential explanatory variables in the data (big model), including PROMIS-GH10 responses, age, and sex.

### 3.4.1 | Classification and regression trees analysis (CART)

Generating a CART model involves selecting explanatory variables, and the split points on those variables, until an optimal tree is constructed. A tree is a prediction algorithm that splits the data at nodes and grows. At each node, the value of one of the explanatory variables (e.g., age >50 or age =<50) determines the next split.

Classification trees and regression trees are adopted when the dependent variable is discrete and continuous, respectively. The selection of explanatory variables and the splits are chosen by minimizing a cost function. While in econometric techniques, the inclusion of explanatory variables (PROMIS-GH10 questions) or their interactions are predefined, CART has the flexibility to include variables and their interactions automatically. For example, the interaction between pain intensity (question 10 of PROMIS-GH10) and other PROMIS-GH10 questions (physical function, fatigue, emotional distress, etc.) may impact the EQ-5D-5L utility values. In addition to accommodating interactions, CART produces algorithms that can readily be expressed and easily understood (Breiman et al., 1984), making it more favorable for mapping.

In direct mapping using CART, the regression trees were generated for EQ-5D-5L utility values. The MSE between the observed and predicted utility values in each node was used to split the data and grow the tree. A range of restrictions was imposed on the tree construction such as the minimum number of observations in the node before the split ($n = 10$), complexity parameters ($cp = 0.001$), 10-fold cross-validation (xval = 10) and setting the "minisplit" and "maxdepth" at different numbers to control the size of the tree (Breiman et al., 1984). The tree construction was stopped when the cost of adding another split to the tree from the current node was above the value of the parameter *cp*.

For indirect mapping using CART, classification trees were grown for all five dimensions of EQ-5D-5L. In growing classification trees, the Gini index was used as the splitting criterion (Breiman et al., 1984; Varian, 2014; Venables & Ripley, 2002). Similar to regression trees, the fully-grown tree was pruned back to the point where cross-validation error was minimized.

The best sized regression and classification trees were chosen according to the smallest misclassification error within the estimation sample and smallest cross-validation error. In case of classification trees, the predicted responses to each dimension of EQ-5D-5L were combined, and an Australian tariff applied to calculate utilities (Norman et al., 2017). An example of a classification tree is presented in detail in Appendix B.

### 3.4.2 | Random forest and bagging (bagged CART)

The single tree generated by CART is highly susceptible to variance in data. There are ensemble approaches such as random forest and bagging that aim to minimize this variance in the prediction and thus improve predictive accuracy (Friedman et al., 2001). However, the lower variance comes with the cost of reduced interpretability, which makes it less desirable for a mapping exercise. The ensemble approaches were adopted in this study to compare the predictive accuracy of models, although they do not generate an algorithm. With these techniques a multitude of decision trees are generated and then aggregated to a single tree based on either the mode (for classification trees) or the mean prediction (for regression trees) of the individual trees (Strobl et al., 2009). Bagging improves variance by averaging the outcome from multiple fully-grown trees on variants of the training data. This reduces the risk of overfitting and substantially improves predictive accuracy compared to a single decision tree (Breiman, 1996, 2001; Liaw & Wiener, 2002). The random forest technique is a modification of the bagging technique. It improves variance by reducing correlation between trees by allowing a selection of a random subset of the explanatory variables at each split to grow independent trees, overcoming the problem of tree correlation inherent in bagging (Boehmke & Greenwell, 2019).

In direct mapping, random forests were developed for EQ-5D-5L utility values by splitting each node using a subset of explanatory variables (PROMIS-GH10 responses, age, and sex) each time. This technique was used to generate 500 decision trees from the randomly selected subsets of the training dataset for each tree. As each tree is well-fitted to a sub-sample of data, the final random forests generated by aggregating these individual trees are expected to fit the whole dataset perfectly. Bagging is performed similarly, however when splitting a node the whole set of explanatory variables is considered.

For indirect mapping using ensemble methods, the aggregated trees were generated for each dimension of EQ-5D-5L, and then the predicted values were combined, and an Australian tariff was applied to obtain utilities (Norman et al., 2017).

### 3.4.3 | Neural networks

Another machine learning method adopted was NN. Although the black box nature of NN is not desirable for this study, they were chosen for their prediction superiority and the ability to perform with a relatively small dataset (Fausett, 1994; Shaikhina & Khovanova, 2017). Moreover, the ability of NN to learn hidden relationships in the data without imposing any fixed relationships makes it an excellent technique for prediction (Tu, 1996).

To estimate utilities with NN, we used a series of multi-layer perceptron feedforward NN, where the information flows from the input nodes (explanatory variables) through the hidden nodes (if any) to the output node (utilities). The model consisted of

an input layer of PROMIS-GH10 items, age and sex (12 nodes), different layers of hidden nodes, and one output node. With direct mapping the output was the EQ-5D-5L utilities, and with indirect mapping, the output was each dimension of EQ-5D-5L.

We also adopted another NN-based technique, QRNN, a mixed technique with the combined advantage of quantile regression and NN. This technique has the ability to model data with non-homogeneous variances and can capture non-linear patterns by using NN, thus advances the standard quantile regression (Cannon, 2011). Moreover, being more resistant to outliers, this technique allowed the predictions to preserve some aspects of the overall distribution of utilities. With this technique we used median regression NN, which was only adopted in direct mapping using PROMIS-GH10 items, age, and sex as inputs nodes to predict EQ-5D-5L utilities (output).

#### 3.4.4 | Least absolute shrinkage and selection operator (LASSO)

We also included the machine learning technique, LASSO, because of its superiority in predicting utilities and model selection. Least absolute shrinkage and selection operator is a type of regression that uses the "shrinkage" technique by imposing a constraint on the parameters that cause regression coefficients for less important variables to shrink toward zero (Tibshirani, 1996). The remaining variables with non-zero coefficients are most strongly associated with the dependent variables, thus enhancing the prediction accuracy and interpretability of the results while reducing the issue of overfitting with regression models. The variable selection feature of LASSO is desirable for mapping. However, in this study we have a relatively small number of explanatory variables. In other mapping exercises using a source measure with a higher number of items and levels a method superior in variable selection could be more beneficial. For the present analysis, we used LASSO for both prediction and variable selection. The former was used as an additional machine learning technique for mapping and the latter was used to enhance the model performance when estimating the hybrid models (see Section 4.2.4).

For direct mapping, LASSO was implemented with several model specifications and the Poisson model was found to perform the best (Park & Hastie, 2007). For prediction with LASSO, two model specifications were considered. The first included only PROMIS-GH10 items, age, age squared, sex, and the second additionally included all two-way interactions of these variables. The training data was used to estimate the model parameters and then the best model was selected based on the smallest out-of-sample MSE. Similar steps were followed to estimate LASSO in the indirect mapping, with the binomial model chosen to predict each dimension of EQ-5D-5L. However, due to computational difficulties, only one set of variables without their interaction were reported in this case.[3]

## 4 | RESULTS

### 4.1 | Descriptive statistics

The sample used to map from PROMIS-GH10 to the EQ-5D-5L comprised of 2015 Australian respondents who completed both instruments. Table 1 provides the sample descriptive statistics.

A high degree of overlap between the source and target measures contributes to more accurate mapping algorithms (Longworth & Rowen, 2013). The overlap between PROMIS-GH10 questions and EQ-5D-5L dimensions and utilities were measured by their correlation, using Spearman's rank correlation coefficients (Zar, 1972). Moderately strong statistically significant correlations between EQ-5D-5L utilities and PROMIS physical (Spearman's rho ($\rho$) = −0.69, $p$ = 0.00) and mental health scores (Spearman's rho ($\rho$) = −0.47, $p$ = 0.00) were observed. These correlations are desirable as the accuracy of a mapping technique depends on the magnitude of overlap between the source and target measures (Longworth & Rowen, 2013).

### 4.2 | Model performance

Table 2 presents the performance of all the econometric and machine learning techniques across a range of criteria. Figures 1–3 compares the distribution of predictions with the observed distribution for the econometric techniques, direct mapping using machine learning techniques, and indirect mapping using machine learning techniques, respectively. Each econometric model was estimated separately for the four sets of covariates described in Section 3.1. We first evaluate the performance of the two types of techniques individually and then compare the two types.

**TABLE 1** Descriptive statistics

| Variables | General population survey |
|---|---|
| Age (years) | |
| Mean (SD) | 48.31 (17.79) |
| Range | 18–89 |
| Female (%) | 53.40% |
| EQ-5D-5L utilities | |
| Mean (SD) | 0.82 (0.25) |
| Range | −0.43 to 1 |
| Utilities <0 (%) | 38 (1.89%) |
| Utilities = 1 (%) | 440 (21.84%) |
| Utilities >0.9 (%) | 1120 (55.58%) |
| PROMIS-GH10 | |
| Physical score (SD) | 14.21 (2.87) |
| Mental score (SD) | 13.22 (3.45) |
| No. of observations | 2015 |

Abbreviation: SD, standard deviation.

### 4.2.1 | Econometric techniques

Models using set 3 and set 4 consistently performed better than those using set 1 and set 2, which is expected since the ordinal nature of PROMIS-GH10 responses were not considered in the latter two sets. Our comparison is therefore based on these two sets of results. Overall, models using set 3 performed better than those using set 4, suggesting a quadratic functional form fits better than the dummy coded age variable. This is also expected since the categorization of a continuous variable would often lose information.

In estimation with ALDVMM[4] we considered two and three component models[5] and found the performance of the former to be superior to the latter. Assuming constant probabilities of component membership, no variables were included in the probabilities of component membership, this might have affected the performance of three component model. The convergence was not achieved when we tried to fit a four-component ALDVMM. As expected, linear regression, median regression and CLAD overpredicted utilities (utilities >1) since they do not consider them being bounded.

Our primary measure of predictive accuracy, the MAE (after truncation), was the lowest for mixture models (ALDVMM and Betamix) using set 3, with values of 0.095826 and 0.096645, respectively. This is consistent with the literature, which suggests their superiority over traditional econometric models for their high level of flexibility (Gray & Hernandez-Alava, 2018; Hernandez-Alava & Wailoo, 2015). The next best performing model was median regression using set 3, whose MAE (after truncation) was 0.099122. The remaining models using set 3 had similar MAEs (after truncation), ranging from 0.103923 to 0.107047. The indirect model GLOGIT performed rather poorly with a MAE of 0.107066. This performance ranking remained overall the same for MSE (after truncation) with the exception that the performance of GLOGIT was not the worst in this case.

Specifically, of all the econometrics techniques, the two mixture models (Betamix and ALDVMM) using set 3 and 4 were the most accurate in predicting the observed mean, with ALDVMM predicting it closer to the observed mean (0.820902). GLOGIT had the poorest performance in predicting the observed mean. All the models performed poorly in terms of predicting the observed minimum utility of −0.426200, with the best performing ones being median regression using set 3 (−0.410107) and two mixture models using set 3 (−0.353980 and −0.367103), and the worst performing being Tobit (−0.242088). The models performed better in terms of predicting the observed maximum utility of one. Among the models whose predictions did not exceed the bound, as expected, the indirect mapping approach GLOGIT performed the best with a maximum utility of one. The others performed similarly, however; the maximum utility ranged from 0.986097 to 0.988465. The comparison between the sample distribution and the distribution of the predictions is more revealing (Figure 1). All the models, apart from the two mixture models, performed poorly toward the extremes. It is particularly interesting that, while median regression could be rated as good as the two mixture models based on Table 2, it is clearly inferior in fitting the different parts of the full distribution.

Based on these comparisons, ALDVMM, closely followed by Betamix, was the best performing among the econometric techniques. The indirect model GLOGIT performed the worst, especially given its relatively poor mean utility prediction.

**TABLE 2** Predicted statistics summary mapping PROMIS-GH10 to EQ-5D-5L

| Models | MAE | | MSE | | Mean (after truncation) | Minimum | Maximum (before truncation) | Maximum (after truncation) | % Of observations predicted >1 before truncation |
|---|---|---|---|---|---|---|---|---|---|
| | Before truncation | After truncation | Before truncation | After truncation | | | | | |
| Actual | | | | | 0.820901 | −0.426230 | 1 | 1 | |
| Econometric models, direct mapping | | | | | | | | | |
| Explanatory variable set 1 | | | | | | | | | |
| Linear regression | 0.142246 | 0.137432 | 0.044354 | 0.042543 | 0.832334 | 0.267356 | 1.198189 | 1 | 5.11% |
| Tobit | 0.131474 | 0.131474 | 0.040423 | 0.040423 | 0.829745 | 0.203345 | 0.964564 | 0.964564 | |
| Median regression | 0.126732 | 0.126144 | 0.040256 | 0.039332 | 0.838288 | 0.181222 | 1.076223 | 1 | 5.26% |
| GLM | 0.135323 | 0.135323 | 0.042167 | 0.042167 | 0.839760 | 0.325465 | 0.985476 | 0.985476 | |
| CLAD | 0.139422 | 0.136421 | 0.044545 | 0.042567 | 0.835377 | 0.173245 | 1.377532 | 1 | 5.46% |
| Betamix | 0.137780 | 0.137780 | 0.040465 | 0.040465 | 0.830632 | 0.123434 | 0.956323 | 0.956323 | |
| ALDVMM | 0.135323 | 0.135323 | 0.038232 | 0.038232 | 0.830053 | 0.111389 | 0.968134 | 0.968134 | |
| Explanatory variable set 2 | | | | | | | | | |
| Linear regression | 0.138100 | 0.130477 | 0.043210 | 0.042005 | 0.832442 | 0.253322 | 1.143212 | 1 | 5.06% |
| Tobit | 0.126243 | 0.126243 | 0.042901 | 0.042901 | 0.833654 | 0.196564 | 0.973114 | 0.973114 | |
| Median regression | 0.125325 | 0.124466 | 0.042132 | 0.038965 | 0.835564 | 0.186231 | 1.032231 | 1 | 5.11% |
| GLM | 0.129445 | 0.129445 | 0.041543 | 0.041543 | 0.834412 | 0.294223 | 0.985234 | 0.985234 | |
| CLAD | 0.135165 | 0.129321 | 0.044532 | 0.041345 | 0.834117 | 0.165564 | 1.144556 | 1 | 5.16% |
| Betamix | 0.121943 | 0.121943 | 0.037553 | 0.037553 | 0.830987 | 0.117326 | 0.975344 | 0.975344 | |
| ALDVMM | 0.120387 | 0.120387 | 0.036890 | 0.036890 | 0.829922 | 0.116745 | 0.977111 | 0.977111 | |
| Explanatory variable set 3 | | | | | | | | | |
| Linear regression | 0.105861 | 0.105061 | 0.034974 | 0.034195 | 0.819438 | −0.283661 | 1.013015 | 1 | 4.96% |
| Tobit | 0.103923 | 0.103923 | 0.030912 | 0.030912 | 0.817443 | −0.243432 | 0.986097 | 0.986097 | |
| Median regression | 0.101734 | 0.099122 | 0.029041 | 0.028455 | 0.829874 | −0.410107 | 1.020771 | 1 | 5.01% |
| GLM | 0.106531 | 0.106531 | 0.031326 | 0.031326 | 0.817477 | −0.296354 | 0.988065 | 0.988065 | |
| CLAD | 0.108825 | 0.107047 | 0.035533 | 0.033462 | 0.829588 | −0.333890 | 1.030432 | 1 | 5.31% |
| Betamix | 0.096645 | 0.096645 | 0.026508 | 0.026508 | 0.820799 | −0.353980 | 0.988395 | 0.988395 | |
| ALDVMM | 0.095826 | 0.095826 | 0.025877 | 0.025877 | 0.820902 | −0.367103 | 0.988465 | 0.988465 | |
| Explanatory variable set 4 | | | | | | | | | |
| Linear regression | 0.109855 | 0.107442 | 0.036302 | 0.035441 | 0.820402 | −0.285332 | 1.039458 | 1 | 5.01% |
| Tobit | 0.105336 | 0.105336 | 0.033271 | 0.033271 | 0.814437 | −0.242088 | 0.986098 | 0.986098 | |
| Median regression | 0.103902 | 0.101391 | 0.031441 | 0.030102 | 0.830179 | −0.375063 | 1.021416 | 1 | 5.21% |
| GLM | 0.107401 | 0.107401 | 0.032052 | 0.032052 | 0.816418 | −0.287088 | 0.988033 | 0.988033 | |
| CLAD | 0.110184 | 0.108371 | 0.035336 | 0.034298 | 0.829330 | −0.318164 | 1.089408 | 1 | 5.26% |
| Betamix | 0.100066 | 0.100066 | 0.029044 | 0.029044 | 0.819360 | −0.355600 | 0.988022 | 0.988022 | |
| ALDVMM | 0.987012 | 0.987012 | 0.027421 | 0.027421 | 0.819057 | −0.366211 | 0.988195 | 0.988195 | |

TABLE 2 (Continued)

| Models | MAE | | MSE | | Mean (after truncation) | Minimum | Maximum (before truncation) | Maximum (after truncation) | % Of observations predicted >1 before truncation |
|---|---|---|---|---|---|---|---|---|---|
| | Before truncation | After truncation | Before truncation | After truncation | | | | | |
| Machine learning, direct mapping | | | | | | | | | |
| CART (regression trees) | 0.126756 | 0.126756 | 0.048433 | 0.048433 | 0.812054 | −0.111331 | 0.981242 | 0.981242 | |
| Random forests | 0.111418 | 0.111418 | 0.037371 | 0.037371 | 0.818166 | −0.202419 | 0.998012 | 0.998012 | |
| Bagged CART | 0.112339 | 0.112339 | 0.041446 | 0.041446 | 0.817192 | −0.196299 | 0.991002 | 0.991002 | |
| NN | 0.107195 | 0.107195 | 0.033278 | 0.033278 | 0.818389 | −0.245290 | 0.992866 | 0.992866 | |
| QRNN | 0.104027 | 0.104027 | 0.031190 | 0.031190 | 0.819744 | −0.300812 | 0.997521 | 0.997521 | |
| LASSO 1 | 0.095523 | 0.095523 | 0.025323 | 0.025323 | 0.820901 | −0.399345 | 0.998733 | 0.998733 | |
| LASSO 2 | 0.101939 | 0.101939 | 0.029339 | 0.029339 | 0.810058 | −0.432911 | 0.964977 | 0.964977 | |
| Econometric models, indirect mapping | | | | | | | | | |
| GLOGIT | 0.107066 | 0.107066 | 0.029267 | 0.029267 | 0.836044 | −0.281108 | 1 | 1 | |
| Machine learning, indirect mapping | | | | | | | | | |
| CART (classification trees) | 0.118269 | 0.118269 | 0.041493 | 0.041493 | 0.860133 | −0.190286 | 1 | 1 | |
| Random forests | 0.107251 | 0.107251 | 0.031279 | 0.031279 | 0.843662 | −0.235079 | 1 | 1 | |
| Bagged CART | 0.111491 | 0.111491 | 0.032466 | 0.032466 | 0.846118 | −0.222931 | 1 | 1 | |
| NN | 0.104729 | 0.104729 | 0.030422 | 0.030422 | 0.831362 | −0.260450 | 1 | 1 | |
| LASSO 1 | 0.104419 | 0.104419 | 0.030680 | 0.030680 | 0.830096 | −0.355210 | 1 | 1 | |

*Note*: Results were obtained from 10-fold cross-validation. Explanatory variables for set 1: the physical and mental health summary scores of PROMIS-GH10 (as continuous variables), age, age squared, sex; set 2: the PROMIS-GH10 items, age, age squared, sex; set 3: the PROMIS-GH10 (as categorical variables), age, age squared, and sex; set 4: the PROMIS-GH10, age, and sex all as categorical variables. LASSO 1: LASSO technique is used for prediction. Explanatory variables (without interactions) are only considered. LASSO 2: LASSO technique is used for prediction. Explanatory variables and their two-way interactions are considered.

Abbreviations: ALDVMM, adjusted limited dependent variable mixture model; Betamix, mixture beta regression model; CLAD, censored least absolute deviation; GLM, generalized linear model; GLOGIT, generalized logistic regression; LASSO, least absolute shrinkage and selection operator; MAE, mean absolute error; MSE, mean squared error; NN, neural networks; PROMIS-GH10, PROMIS short form Global Health 10; QRNN, quantile (median) regression neural networks.

### 4.2.2 | Machine learning

None of the machine learning techniques over predicted the utilities, thus the MAE before truncation and after truncation were identical. We used the LASSO technique to estimate a model without variables' interaction (LASSO 1), and a model with variables' interaction (LASSO 2). While LASSO 1 was used in both direct and indirect mapping, the inclusion of interactions was specific to direct mapping due to computational difficulty. The direct LASSO 1 model performed the best with a MAE of 0.095523 while regression trees the worst with a MAE of 0.126756.

The best performing machine learning technique, direct LASSO 1, selected following variables to optimize the prediction: PROMIS-GH10 questions 1 (general health), question 4 (mental health), question 5 (social activities), question 6 (physical activities), question 7 (pain), question 9 (social activities), question 10 (emotional problems), sex, age, and age squared. The inclusion of two-way interactions in the direct LASSO 2 model, somewhat surprisingly, worsened the predictive performance and the MAE increased from 0.095523 to 0.101939.

In comparison, PROMIS-GH10 question 6 (physical activities), question 7 (pain), question 9 (social activities) and question 10 (emotional problems) had high importance in predicting utilities with the regression tree as most splits for growing the tree

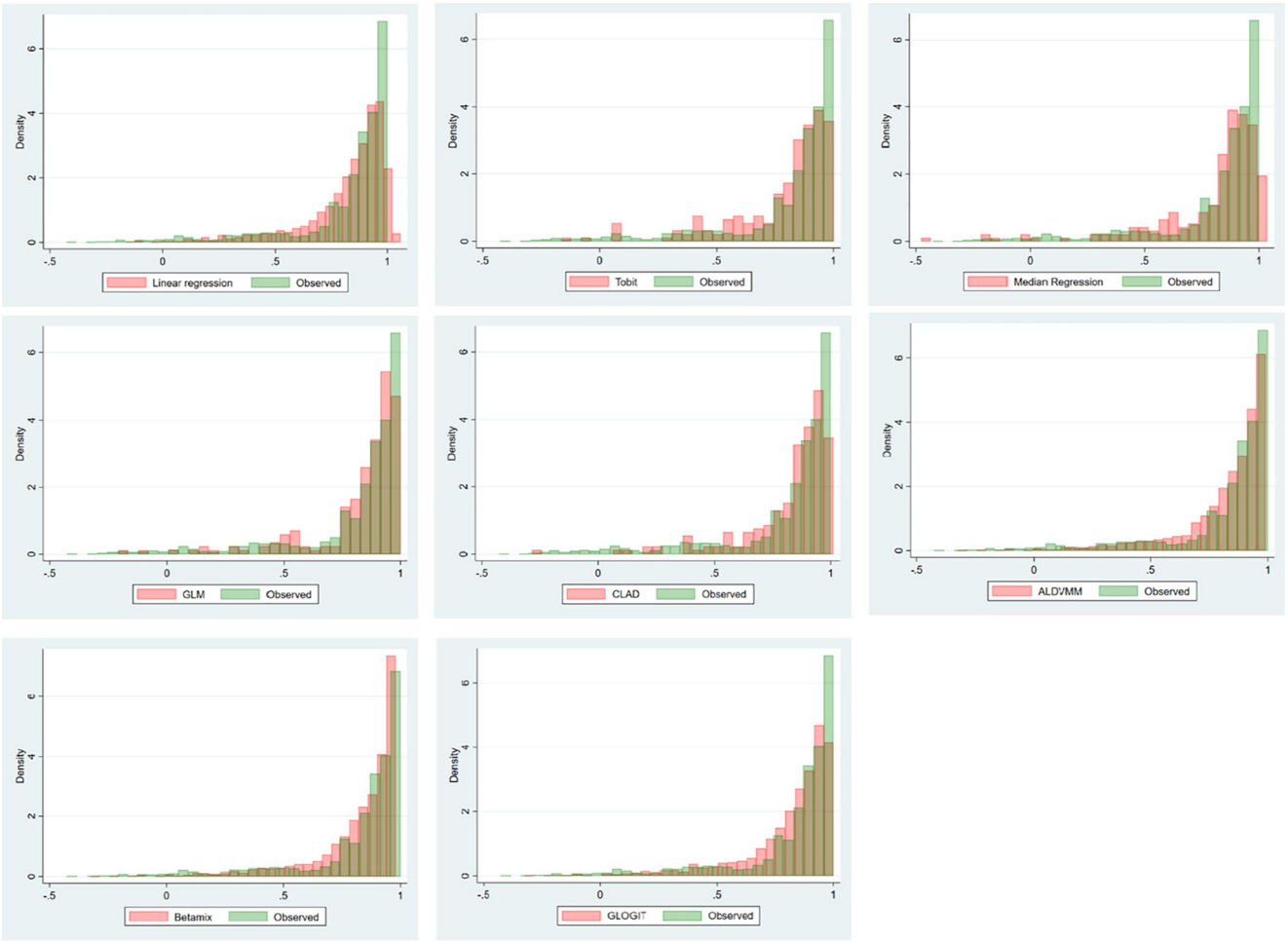


**FIGURE 1** Distribution of the observed versus predicted utilities using the econometric techniques. ALDVMM, adjusted limited dependent variable mixture model; Betamix, mixture beta regression model; CLAD, censored least absolute deviation; GLM, generalized linear model; GLOGIT, generalized logistic regression; MR, median regression [Colour figure can be viewed at wileyonlinelibrary.com]

were based on responses to these questions. In classification trees, a set of different variables had high importance depending on the dimension. As an example, in predicting the depression and anxiety dimension, PROMIS-GH10 questions 10 (emotional problems) and question 4 (mental health states) had a greater contribution.

The predictive accuracy of regression and classification trees improved when random forests and bagging were applied. For regression trees, the MAE improved from 0.126756 to 0.111418 with random forests and to 0.112339 with bagging. For classification trees, the MAE improved from 0.118269 to 0.107251 with random forests and to 0.111491 with bagging. An example of classification tree prediction is presented in Appendix B.

Direct mapping with NN further improved the MAE to 0.107195 and indirect mapping with NN resulted in the MAE of 0.104729. With QRNN, the MAE again improved to 0.104027.

This performance ranking remained the same for MSE. The direct machine learning techniques were more accurate in predicting the observed mean compared to the indirect ones, and direct LASSO 1 was the most accurate in predicting 0.820901 exactly. Classification trees were the worst, predicting the mean as 0.860133. Both direct and indirect LASSO 1 and direct LASSO 2 performed well in terms of predicting the observed minimum utility of −0.426230, with the best performing ones being direct LASSO 2 (−0.432911) and direct LASSO 1 (−0.399345). Apart from direct LASSO 2, all the machine learning techniques performed well in predicting the maximum utilities, with the indirect techniques predicting the exact maximum utility value of one, and the direct techniques predicting the maximum in the range 0.981242–0.998733. Direct LASSO 2 predicted 0.9649770.

The comparison between the sample distribution and the distribution of the predictions gives more insights on how these techniques performed. Figures 2 and 3 suggest that, except LASSO models, all direct techniques fitted the distribution poorly. The indirect approaches using CART, random forest, bagged CART, and NN fitted better than the direct approaches.


**FIGURE 2** Distribution of the observed versus predicted utilities using direct mapping with machine learning techniques. LASSO 1: LASSO technique is used for prediction. Explanatory variables (without interactions) are only considered. LASSO 2: LASSO technique is used for prediction. Explanatory variables and their two-way interactions are considered. LASSO, least absolute shrinkage and selection operator; NN, neural networks; QRNN, quantile (median) regression neural networks [Colour figure can be viewed at wileyonlinelibrary.com]

Overall, the direct LASSO 1 seemed to dominate all the other machine learning techniques. The indirect LASSO 1 also performed the best among all the indirect techniques. CART techniques (classification trees and regression trees) appeared to perform the worst overall. The performance of indirect approaches in each dimension of EQ-5D-5L is presented in Appendix D.

### 4.2.3 | Comparison of econometric and machine learning techniques

The direct LASSO 1 out-performed the best performing econometric model (ALDVMM using set 3) for all criteria, although only by a relatively small margin. The former had slightly smaller MAE and MSE, with the minimum prediction closer to the observed and the maximum better (i.e., closer to one). Both were able to predict the observed mean and fitted the distribution similarly. It is worth mentioning that, within the indirect mapping approaches, LASSO 1 out-performed the best performing econometric technique (GLOGIT).

### 4.2.4 | Estimating the hybrid model

Overall, the LASSO and ALDVMM techniques out-performed all the other machine learning and econometric techniques, respectively. However, the calculation of standard errors and variance-covariance matrices with LASSO is not straightforward. Moreover, LASSO regularization excludes some variables to estimate a simpler model. The correlation between the selected

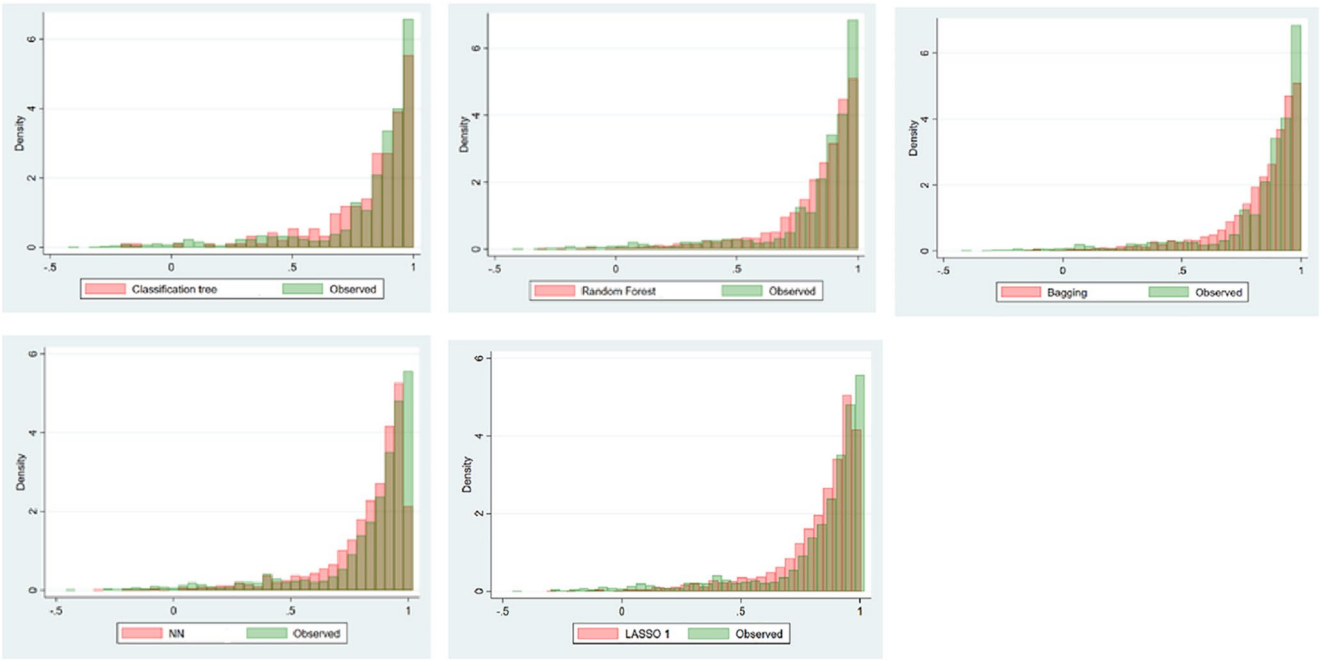


**FIGURE 3** Distribution of the observed versus predicted utilities using indirect mapping with machine learning techniques. LASSO 1: LASSO technique is used for prediction. Explanatory variables (without interactions) are only considered. LASSO, least absolute shrinkage and selection operator; NN, neural networks; QRNN, quantile (median) regression neural networks [Colour figure can be viewed at wileyonlinelibrary.com]

variables and those excluded might lead to bias in the estimated coefficients (Ahrens et al., 2019; Barlin et al., 2013; Lee et al., 2016).

Overcoming the limitation of LASSO and making use of the variable selection feature of this technique, we developed additional hybrid models (Hybrid 1 and Hybrid 2) wherein we improved model performance by combining machine learning and econometric techniques. Specifically, we first selected the variables using LASSO and then re-estimated ALDVMM using these variables.

However, choosing the variables using the whole sample and then re-fitting the model can be problematic as only the significant variables are chosen, and the standard errors cannot be trusted (Lee et al., 2016; Mullainathan & Spiess, 2017). One way to address this issue is to divide the dataset into two sub-samples and use one for the variable selection, and the other for estimation of the models (Zhao et al., 2017). Following this approach, we used half of the data for LASSO variable selection and then re-fitted ALDVMM with the selected variables using the other half of the data. Also, we estimated the ALDVMM with explanatory variable set 3 and LASSO separately with the exact estimation and validation sample (50% of the sample) to be able to compare the results.

For the Hybrid 1 model, we used the LASSO technique for variable selection among PROMIS-GH10 items, age, and sex (without their two-way interactions). For the Hybrid 2 model, we additionally included two-way interactions for variable selection.

The results presented in Table 3 suggests the Hybrid 1 model resulted in improved utility predictions in the extremes, with the MAE lower than the ALDVMM. The selected variables enabled ALDVMM to predict the minimum utility of −0.322932, an improvement of 0.016082 compared to −0.306850 predicted by the ALDVMM with set 3. Moreover, the superiority of mixture models in accommodating multimodality combined with the selected variables, resulted in better accuracy in predicting full health utilities. However, ALDVMM performed slightly better in predicting the exact mean. These results suggest LASSO's variable selection feature resulted in improving the performance of the ALDVMM in terms of MAE, minimum, and maximum utilities.

The hybrid model with the inclusion of interactions (Hybrid 2), on the other hand, did not improve the overall performance of the model. Although this model was superior in predicting the minimum utility, it was at the cost of a more inaccurate prediction of the mean and maximum utility.

Overall, these results suggest that not only does direct LASSO one out-perform all other models in prediction, but also utilizing LASSO's variable selection feature improved ALDVMM's predictive performance.

**TABLE 3** Performance of hybrid models

| Models | MAE | Rank in MAE | MSE | Rank in MSE | Mean | Rank in mean | Minimum | Rank in minimum | Maximum | Rank in maximum |
|---|---|---|---|---|---|---|---|---|---|---|
| Hybrid 1 | 0.096125 | 2 | 0.026310 | 3 | 0.826410 | 3 | −0.322932 | 3 | 0.998757 | 2 |
| Hybrid 2 | 0.098943 | 4 | 0.298421 | 4 | 0.815864 | 4 | −0.405733 | 1 | 0.979543 | 4 |
| LASSO 1[a] | 0.095993 | 1 | 0.025773 | 1 | 0.826159 | 1 | −0.347765 | 2 | 0.998831 | 1 |
| LASSO 2[a] | 0.995188 | 5 | 0.029542 | 5 | 0.810641 | 5 | −0.449753 | 5 | 0.969521 | 5 |
| ALDVMM[a] | 0.096341 | 3 | 0.026052 | 2 | 0.826335 | 2 | −0.306850 | 4 | 0.988367 | 3 |
| Actual observations in the validation sample (50% of dataset) | | | | | 0.826099 | | −0.426230 | | 1 | |

*Note*: Hybrid 1: explanatory variables (without interactions) are selected by LASSO and the ALDVMM is re-estimated with the selected variables. Hybrid 2: explanatory variables (variables and their two-way interactions) are selected by LASSO and the ALDVMM is re-estimated with the selected variables.

Abbreviations: ALDVMM, adjusted limited dependent variable mixture model; LASSO, least absolute shrinkage and selection operator.

[a]These three models are re-estimated using 50% estimation and 50% validation sample to be comparable with Hybrid models, thus the statistics are different from ones previously reported in Table 2.

## 5 | DISCUSSION

There has been increased interest in machine learning techniques in the health economics literature with the presumption they will out-perform standard econometric techniques (Konig et al., 2013; Kreif et al., 2015; Schilling et al., 2017). However, there has been a realization that while econometric techniques can perform poorly regarding predicting complex and non-linear relationships, they are easier to implement and are superior in explaining and interpreting those relationships. This has inspired the use of hybrid econometric-machine learning techniques to predict and interpret complex relations (Boelaert & Ollion, 2018; Böheim & Stöllinger, 2021; Kauffman et al., 2017; Malhotra, 2021; Yu et al., 2007; Zheng et al., 2017).

This paper explored the feasibility of using machine learning techniques and combining them with econometric methods as a valuable tool for mapping PROMs to MAUIs. We used machine learning techniques to map from PROMIS-GH10 to EQ-5D-5L and compared their performance to the standard econometric techniques previously adopted in the literature. Both direct and indirect techniques of mapping were applied, and utilities were estimated for six machine learning techniques (CART, random forests, bagged CART, NN, QRNN, and LASSO) and eight econometric techniques (linear regression, Tobit, GLM, median regression, CLAD, Betamix, ALDVMM, and GLOGIT).

The direct LASSO 1 model performed the best across the range of econometric and machine learning techniques, followed by ALDVMM with MAEs of 0.095523 and 0.095826, respectively. Similar to those observed in a previous study mapping PROMIS-GH10 to EQ-5D-3L by Thompson et al. (2017) using a substantially larger sample ($n = 13{,}955$) with MAEs ranging between 0.069 and 0.144.

CART techniques (classification trees and regression trees) were the worst performing machine learning techniques. Consistent with the literature, applying ensemble algorithms (Random Forest and Bagging) to them is essential as it increases prediction accuracy, although this improved performance comes at the cost of interpretability (Breiman, 1996, 2001; Friedman et al., 2001; Liaw & Wiener, 2002).The mapping literature has been dominated by the efforts of selecting optimal model specifications while less attention has been paid to variable selection. Our results suggest that the latter is equally important and should be considered in mapping exercises. Traditionally variables have been selected using a "cherry picking" approach or a "kitchen sink" approach, where the former is based on theory and the latter relies on the implicit variable selection through the coefficient values (Chen et al., 2019). The advantage of using machine learning techniques for variable selection has been emphasized in the literature (Athey & Imbens, 2019). However, the value of using LASSO for variable selection continues to be debated, with recent studies comparing the performance of several techniques reporting mixed results (Vasquez et al., 2016; Zou, 2006).

While LASSO out-performed the other techniques in prediction, the calculation of standard errors and variance-covariance matrices is not straight forward for LASSO, like any other machine learning techniques. Consequently, if a researcher was interested in more than the deterministic results of a cost-effectiveness analysis (e.g., probabilistic sensitivity analysis) then machine learning techniques could not be used to generate a mapping algorithm. Nevertheless, machine learning techniques' variable selection feature can be adopted to enhance econometric techniques. As it is examined in this study, combining this feature with the best performing econometric techniques resulted in a hybrid model with improved predictive performance in several

criteria. The standard errors and variance-covariance matrices for the hybrid model are easy to obtain. However, to address overfitting bias, and acquire reliable standard errors, the parametric models are required to be estimated on a different sample.

Based on the performance of the hybrid models, we have proposed two algorithms to map from the PROMIS-GH10 to EQ-5D-5L in Appendix C. One is based on ALDVMM with explanatory variable set 3, and the other one is based on the Hybrid 1 model which includes variables selected by LASSO technique and it is re-estimated with ALDVMM. The corresponding variance-covariance matrices are also presented in Appendix E.

It should be noted that in our estimation, the LASSO variable selection was implemented for several model specifications, with Poisson performing the best for direct mapping and binomial for the indirect mapping. However, the ALDVMM model was not included in the comparison as such an algorithm has yet to be developed. Our hybrid model represents a pragmatic approach that combines the power of LASSO variable selection and the flexibility of ALDVMM model specification. Indeed, this approach improved the original ALDVMM (without variable selection) on almost every metric. Nevertheless, how to implement the LASSO variable selection within the ALDVMM model and whether this may further improve the predictive performance are interesting questions and should be explored in future research.

In prediction with LASSO and the hybrid model, the inclusion of two-way interactions led to worse predictive performance than the exclusion of the interactions. This may be due to the high correlations between predictors and the relatively small sample size (so the interactions cannot be precisely estimated). However, it should be noted that considering interactions in LASSO led to the best performance on predicting lower utilities (but very poor performance on predicting high utility values), suggesting that when the health is poor, the interaction may play a more important role in predicting the utilities.

We adopted a sample splitting approach to obtain reliable standard errors to address regularization bias associated with the LASSO (Mullainathan & Spiess, 2017). However, there might be some concerns around the randomness associated with the method (i.e., different splits would yield different results). One possible way to resolve this issue is to perform multiple random splits and aggregate the information accordingly (Meinshausen et al., 2009). This should be explored in future studies.

Given these limitations with machine learning techniques for variable selection in general, including LASSO, these techniques should be used and interpreted cautiously. However, we recommend a hybrid model can be regarded as a supplementary tool in mapping exercises to guide the variable selection and maximize predictive performance.

While an advantage of a machine learning technique is its capability to learn and improve its performance (Breiman et al., 1984), model interpretability and explainability restrict its application to mapping. The "black box" nature of some of the machine learning techniques imposes a significant limitation on their adoption as there is no algorithm that can be reported for another researcher to use. However, as shown in this paper, certain machine learning techniques like LASSO alongside standard econometric mapping techniques can enhance predictions by improving variable selection.

Moreover, the emerging field in machine learning of explainable Artificial Intelligence (AI) has demonstrated practical success in providing an insight into the "black box" (Holzinger et al., 2017). We believe research in explainable AI could facilitate the implementation of machine learning in patient reported data, and specifically in mapping.

Moreover, some machine learning techniques can optimize a joint loss function comprised of different items without collapsing them into an overall score. Recent machine learning literature has attempted to address this by relaxing the hypothesis of the piecewise linear loss function in adopting multi-task learning (Brault et al., 2019; Dosovitskiy & Djolonga, 2019; Shoshan et al., 2019; Wang et al., 2019). Brault et al. (2019) proposed Infinite Task Learning, which jointly solves parametrized tasks for a continuum of parameters. Dosovitskiy and Djolonga (2019) proposed an approach "you only train once (YOTO)", which trains one models across the entire space of different loss weightings. However, evidence on the reliable performance of these models for relatively small datasets like ours is not sufficiently validated and should be explored in the future studies.

While the advantage of using a machine learning technique is its capability to learn and improve its performance (Breiman et al., 1984), this capability is limited by the availability of data. Machine learning is data driven and usually requires a large dataset (optimally 75–100 observations per class) to work efficiently. In comparison, this study had 2015 observations, with some levels having less than 10 observations. This was more pronounced in lower utilities as only 2% of the respondents in the full sample reported negative utilities. While this is a smaller sample than some machine learning studies in other disciplines, ours is a substantial sample relative to other studies using machine learning in patient level health outcome research (Konig et al., 2013; Kreif et al., 2015; Schilling et al., 2017). With the likelihood of patient level datasets being relatively smaller in most future studies, we believe our analysis offer important insights for future studies aimed at evaluating a range of methodological techniques for mapping.

Our analysis was based on a single case study of mapping from PROMIS-GH10 to EQ-5D-5L. As with all mapping studies, there is uncertainty around the results and the differences in MAE, our primary measure of model performance. In fact, for a different dataset another model could perform better. Thus, future research applying machine learning to other data sets, involving different instruments, sample sizes, and types of respondents, would be needed to further validate our results.

## 6 | CONCLUSION

This study makes two significant contributions to the literature. This is the first study to simultaneously consider a broad range of econometric and machine learning techniques for mapping and to compare their performance in predicting utilities. While most mapping literature has exclusively used econometric techniques that are parametric in nature and require some tweaking (e.g., truncation, stepwise regression) that can lead to biases. A key advantage of using machine learning techniques for mapping is that they overcome the need to prespecify the functional specifications of the models. This would be an advantage if the PROM had a high number of items and levels. Our approach of combining econometric and machine learning techniques brings new insights to the mapping literature. Future research on mapping patient outcome data would further validate our results for predictive accuracy of machine learning techniques and hybrid models for different datasets. The second contribution of this study is the development of two mapping algorithms to map from the PROMIS-GH10 to the EQ-5D-5L.

## ACKNOWLEDGMENT

We would like to thank participants at the at 40th Australian Health Economics Society Conference, the eighth Meeting of the International Academy of Health Preference Research in 2018, the UK Health Economics Study Group Summer Meeting in 2019, and seminar participants at the University of York 2019, for their comments and useful insights into the study methods. The authors would also like to acknowledge the feedback from Dr. Gang Chen, Associate Professor in Center for Health Economics (CHE) at Monash University. This project was funded by the Central Coast Local Health District (CCLHD), NSW, Australia. Ethics approval for this study was obtained from the Macquarie University Ethics Committee on November 8, 2017 (Reference number: 5201700985).

Open access publishing facilitated by Macquarie University, as part of the Wiley - Macquarie University agreement via the Council of Australian University Librarians.

## CONFLICT OF INTEREST

All the authors declare that they have no conflict of interest.

## AUTHOR CONTRIBUTION

Mona Aghdaee, Bonny Parkinson, Kompal Sinha, Yuanyuan Gu, Rajan Sharma, Emma Olin and Henry Cutler contributed to the conception and design of this mapping study. Mona Aghdaee conducted the statistical analysis. All the authors contributed to the interpretation of data; drafting the article, revising it critically for the intellectual content and final approval version to be published.

## DATA AVAILABILITY STATEMENT

Research data are not shared.

## ORCID

*Mona Aghdaee* https://orcid.org/0000-0002-2570-1685
*Bonny Parkinson* https://orcid.org/0000-0002-2137-0962
*Kompal Sinha* https://orcid.org/0000-0003-4318-6100
*Yuanyuan Gu* https://orcid.org/0000-0002-3816-9106
*Rajan Sharma* https://orcid.org/0000-0002-4474-5296
*Emma Olin* https://orcid.org/0000-0003-1762-5851
*Henry Cutler* https://orcid.org/0000-0002-6015-092X

## ENDNOTES

[1] For the machine learning techniques, we used the random pre-defined 10 subsamples used in the econometric techniques to predict the utilities. This enabled the comparability between machine learning and econometrics models.

[2] The MAE and MSE are expressed as: $MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$ and $MSE = \frac{\sum_{i=1}^{n} (yi - \hat{y}_i)^2}{n}$ where $n$ is the number of observations, $y_i$ is the observed value of utilities, and $\hat{y}_i$ is predicted value by the mapping algorithm.

[3] The model failed to converge when all interactions were included.

[4] ALDVMM was estimated by using the Stata command "aldvmm".

[5] ALDVMM accounts for multimodality by modeling utilities as a mixture of multiple components, each representing a cluster of respondents with similar utility scores. Here the estimated models consisted of two components.

## REFERENCES


Ahrens, A., Hansen, C. B., & Schaffer, M. (2019). *LASSOPACK: Stata module for lasso, square-root lasso, elastic net, ridge, adaptive lasso estimation and cross-validation*.

Ara, R., & Brazier, J. (2008). Deriving an algorithm to convert the eight mean SF-36 dimension scores into a mean EQ-5D preference-based score from published studies (where patient level data are not available). *Value in Health*, *11*(7), 1131–1143. https://doi.org/10.1111/j.1524-4733.2008.00352.x

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Australian Bureau of Statistics. (2017). *Australian demographic statistics*.

Barlin, J. N., Zhou, Q., St Clair, C. M., Iasonos, A., Soslow, R. A., Alektiar, K. M., Hensley, M. L., Leitao, M. M., Barakat, R. R., & Abu-Rustum, N. R. (2013). Classification and regression tree (CART) analysis of endometrial carcinoma: Seeing the forest for the trees. *Gynecologic Oncology*, *130*(3), 452–456. https://doi.org/10.1016/j.ygyno.2013.06.009

Basu, A., & Manca, A. (2012). Regression estimators for generic health-related quality of life and quality-adjusted life years. *Medical Decision Making*, *32*(1), 56–69. https://doi.org/10.1177/0272989x11416988

Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. CRC Press.

Boelaert, J., & Ollion, É. (2018). The great regression. *Revue Française de Sociologie*, *59*(3), 475–506. https://doi.org/10.3917/rfs.593.0475

Böheim, R., & Stöllinger, P. (2021). Decomposition of the gender wage gap using the LASSO estimator. *Applied Economics Letters*, *28*(10), 817–828. https://doi.org/10.1080/13504851.2020.1782332

Brault, R., Lambert, A., Szabó, Z., Sangnier, M., & d'Alché-Buc, F. (2019). Infinite task learning in RKHSS. In *Paper presented at the 22nd International Conference on Artificial Intelligence and Statistics*.

Brazier, J. E., Yang, Y., Tsuchiya, A., & Rowen, D. L. (2010). A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *The European Journal of Health Economics*, *11*(2), 215–225. https://doi.org/10.1007/s10198-009-0168-z

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth International Group.

Cannon, A. J. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers & Geosciences*, *37*(9), 1277–1284. https://doi.org/10.1016/j.cageo.2010.07.005

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., DeVellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J. S., Pilkonis, P., Revicki, D., … & Hays, R. (2010). The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology*, *63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Chaudhry, S., Jin, L., & Meltzer, D. (2005). Use of a self-report-generated Charlson comorbidity index for predicting mortality. *Medical Care*, *43*(6), 607–615. https://doi.org/10.1097/01.mlr.0000163658.65008.ec

Chen, J. C., Dunn, A., Hood, K., Driessen, A., & Batch, A. (2019). Off to the races: A comparison of machine learning and alternative data for predicting economic indicators. In *Big Data for 21st Century Economic Statistics*. University of Chicago Press.

Crott, R., & Briggs, A. (2010). Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences. *The European Journal of Health Economics*, *11*(4), 427–434. https://doi.org/10.1007/s10198-010-0233-7

Dakin, H., Abel, L., Burns, R., & Yang, Y. (2018). 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*, *16*(1), 1–9. https://doi.org/10.1186/s12955-018-0857-3

Dakin, H., Petrou, S., Haggard, M., Benge, S., & Williamson, I. (2010). Mapping analyses to estimate health utilities based on responses to the OM8-30 Otitis Media Questionnaire. *Quality of Life Research*, *19*(1), 65–80. https://doi.org/10.1007/s11136-009-9558-z

Department of Health. (2016). *Guidelines for preparing a submission to the pharmaceutical benefits advisory committee*. Commonwealth of Australia

Dosovitskiy, A., & Djolonga, J. (2019). *You only train once: Loss-conditional training of deep networks*. Paper presented at the International Conference on Learning Representations.

Evans, J. P., Smith, A., Gibbons, C., Alonso, J., & Valderas, J. M. (2018). The national Institutes of health patient-reported outcomes measurement information system (PROMIS): A view from the UK. *Patient Related Outcome Measures*, *9*, 345–352. https://doi.org/10.2147/PROM.S141378

Fausett, L. V. (1994). *Fundamentals of neural networks: Architectures, algorithms and applications*. Prentice Hall.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics.

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), 137–146. https://doi.org/10.1007/s11222-009-9153-8

Gray, L. A., & Alava, M. H. (2018). A command for fitting mixture regression models for bounded dependent variables using the beta distribution. *STATA Journal*, *18*(1), 51–75. https://doi.org/10.1177/1536867x1801800105

Gray, L. A., Rivero-Arias, O. R., & Clarke, P. M. (2006). Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Medical Decision Making*, *26*(1), 18–29. https://doi.org/10.1177/0272989X05284108

Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Quality of Life Research*, *18*(7), 873–880. https://doi.org/10.1007/s11136-009-9496-9

Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., Bonsel, G., & Badia, X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, *20*(10), 1727–1736. https://doi.org/10.1007/s11136-011-9903-x

Hernandez-Alava, M., & Wailoo, A. (2015). Fitting adjusted limited dependent variable mixture models to EQ-5D. *STATA Journal*, *15*(3), 737–750. https://doi.org/10.1177/1536867x1501500307

Hernandez-Alava, M., Wailoo, A., Wolfe, F., & Michaud, K. (2013). The relationship between EQ-5D, HAQ and pain in patients with rheumatoid arthritis. *Rheumatology*, *52*(2), 944–950. https://doi.org/10.1093/rheumatology/kes400

Hernandez-Alava, M., Wailoo, A., Wolfe, F., & Michaud, K. (2014). A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes. *Medical Decision Making*, *34*(7), 919–930. https://doi.org/10.1177/0272989X13500720

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). *What do we need to build explainable AI systems for the medical domain?* arXiv preprint arXiv.

Janssen, M. F., Birnie, E., Haagsma, J. A., & Bonsel, G. J. (2008). Comparing the standard EQ-5D three-level system with a five-level version. *Value in Health*, *11*(2), 275–284. https://doi.org/10.1111/j.1524-4733.2007.00230.x

Kaambwa, B., Bryan, S., Barton, P., Parker, H., & Martin, G. (2006). *Relationship between the EuroQol-5d and Barthel Index—examining the use of proxy outcome measures for older people*.

Kauffman, R. J., Kim, K., Lee, S.-Y. T., Hoang, A.-P., & Ren, J. (2017). Combining machine-based and econometrics methods for policy analytics insights. *Electronic Commerce Research and Applications*, *25*, 115–140. https://doi.org/10.1016/j.elerap.2017.04.004

Kearns, B., Ara, R., Wailoo, A., Manca, A., Alava, M. H., Abrams, K., & Campbell, M. (2013). Good practice guidelines for the use of statistical regression models in economic evaluations. *PharmacoEconomics*, *31*(8), 643–652. https://doi.org/10.1007/s40273-013-0069-y

Khan, I., & Morris, S. (2014). A non-linear beta-binomial regression model for mapping EORTC QLQ-C30 to the EQ-5D-3L in lung cancer patients: A comparison with existing approaches. *Health and Quality of Life Outcomes*, *12*(1), 163. https://doi.org/10.1186/s12955-014-0163-7

Konig, H. H., Leicht, H., Bickel, H., Fuchs, A., Gensichen, J., Maier, W., Mergenthal, K., Riedel-Heller, S., Schafer, I., Schon, G., Weyerer, S., Wiese, B., van den Bussche, H., Scherer, M., Eckardt, M., & MultiCare study group. (2013). Effects of multiple chronic conditions on health care costs: An analysis based on an advanced tree-based regression model. *BMC Health Services Research*, *13*(1), 219. https://doi.org/10.1186/1472-6963-13-219

Kreif, N., Grieve, R., Diaz, I., & Harrison, D. (2015). Evaluation of the effect of a continuous treatment: A machine learning approach with an application to treatment for traumatic brain injury. *Health Economics*, *24*(9), 1213–1228. https://doi.org/10.1002/hec.3189

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, *44*(3), 907–927. https://doi.org/10.1214/15-aos1371

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. STATA Press.

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

Malhotra, A. (2021). A hybrid econometric–machine learning approach for relative importance analysis: Prioritizing food policy. *Eurasian Economic Review*, *11*(3), 1–33. https://doi.org/10.1007/s40822-021-00170-9

Manning, W. G., & Mullahy, J. (2001). Estimating log models: To transform or not to transform? *Journal of Health Economics*, *20*(4), 461–494. https://doi.org/10.1016/s0167-6296(01)00086-8

Medical Services Advisory Committee. (2016). *Technical guidelines for preparing assessment reports for the Medical Services Advisory Committee –medical service type: Therapeutic version 2.0*. Department of Health, Commonwealth of Australia. http://www.msac.gov.au/internet/msac/publishing.nsf/Content/0BD63667C984FEEACA25801000123AD8/$File/TherapeuticTechnicalGuidelines-Final-March2016-Version2.0-accessible.pdf

Meinshausen, N., Meier, L., & Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, *104*(488), 1671–1681. https://doi.org/10.1198/jasa.2009.tm08647

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *The Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

National Institute for Health and Care Excellence. (2013). *Technology appraisals: A guides to the methods of technology appraisal*.

Nijagal, M. A., Wissig, S., Stowell, C., Olson, E., Amer-Wahlin, I., Bonsel, G., Brooks, A., Coleman, M., Devi Karalasingam, S., Duffy, J. M. N., Flanagan, T., Gebhardt, S., Greene, M. E., Groenendaal, F., R Jeganathan, J. R., Kowaliw, T., Lamain-de-Ruiter, M., Main, E., Owens, M., & Franx, A. (2018). Standardized outcome measures for pregnancy and childbirth, an ICHOM proposal. *BMC Health Services Research*, *18*(1), 953. https://doi.org/10.1186/s12913-018-3732-3

Norman, R., Viney, R., Mulhern, B., Brazier, J. E., Ratcliffe, J., Lancsar, E., & Flattery, M. (2017). A large Australian DCE with duration and dead to value EQ-5D-5L health states. In *Paper presented at the 2017 EuroQol Meeting Barcelona*.

Park, M., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B*, *69*(4), 659–677. https://doi.org/10.1111/j.1467-9868.2007.00607.x

Park, S., & Basu, A. (2018). Alternative evaluation metrics for risk adjustment methods. *Health Economics*, *27*(6), 984–1010. https://doi.org/10.1002/hec.3657

Petrou, S., Rivero-Arias, O., Dakin, H., Longworth, L., Oppe, M., Froud, R., & Gray, A. (2015). Preferred reporting items for studies mapping onto preference-based outcome measures: The MAPS statement. *PharmacoEconomics*, *33*(10), 985–991. https://doi.org/10.1007/s40273-015-0319-2

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, *25*(3), 303–325. https://doi.org/10.1016/0304-4076(84)90004-6

Revicki, D. A., Kawata, A. K., Harnam, N., Chen, W. H., Hays, R. D., & Cella, D. (2009). Predicting EuroQol (EQ-5D) scores from the patient-reported outcomes measurement information system (PROMIS) global items and domain item banks in a United States sample. *Quality of Life Research*, *18*(6), 783–791. https://doi.org/10.1007/s11136-009-9489-8

Rowen, D., Brazier, J., & Roberts, J. (2009). Mapping SF-36 onto the EQ-5D index: How reliable is the relationship? *Health and Quality of Life Outcomes*, *7*(1), 27. https://doi.org/10.1186/1477-7525-7-27

Salinas, J., Sprinkhuizen, S. M., Ackerson, T., Bernhardt, J., Davie, C., George, M. G., Gething, S., Kelly, A. G., Lindsay, P., Liu, L., Martins, S. C., Morgan, L., Norrving, B., Ribbers, G. M., Silver, F. L., Smith, E. E., Williams, L. S., & Schwamm, L. H. (2016). An international standard set of patient-centered outcome measures after stroke. *Stroke*, *47*(1), 180–186. https://doi.org/10.1161/STROKEAHA.115.010898

Schilling, C., Mortimer, D., & Dalziel, K. (2017). Using CART to identify thresholds and hierarchies in the determinants of funding decisions. *Medical Decision Making*, *37*(2), 173–182. https://doi.org/10.1177/0272989X16638846

Schilling, C., Mortimer, D., Dalziel, K., Heeley, E., Chalmers, J., & Clarke, P. (2016). Using Classification and Regression Trees (CART) to identify prescribing thresholds for cardiovascular disease. *PharmacoEconomics*, *34*(2), 195–205. https://doi.org/10.1007/s40273-015-0342-3

Shaikhina, T., & Khovanova, N. A. (2017). Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial Intelligence in Medicine*, *75*, 51–63. https://doi.org/10.1016/j.artmed.2016.12.003

Sharma, R., Gu, Y., Sinha, K., Aghdaee, M., & Parkinson, B. (2019). Mapping the strengths and difficulties questionnaire onto the child health utility 9D in a large study of children. *Quality of Life Research*, *28*(9), 2429–2441. https://doi.org/10.1007/s11136-019-02220-x

Shaw, J. W., Pickard, A. S., Yu, S., Chen, S., Iannacchione, V. G., Johnson, J. A., & Coons, S. J. (2010). A median model for predicting United States population-based EQ-5D health state preferences. *Value in Health*, *13*(2), 278–288. https://doi.org/10.1111/j.1524-4733.2009.00675.x

Shoshan, A., Mechrez, R., & Zelnik-Manor, L. (2019). Dynamic-net: Tuning the objective without re-training for synthesis tasks. In *Paper presented at the IEEE International Conference on Computer Vision*.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Sullivan, P. W., & Ghushchyan, V. (2006). Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. *Medical Decision Making*, *26*(4), 401–409. https://doi.org/10.1177/0272989X06290496

Thompson, C., Sansoni, J., Morris, D., Capell, J., & Williams, K. (2016). *Patient-reported outcome measures: An environmental scan of the Australian healthcare sector*. ACSQHC.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tosh, J. C., Longworth, L. J., & George, E. (2011). Utility values in national institute for health and clinical excellence (NICE) technology appraisals. *Value in Health*, *14*(1), 102–109. https://doi.org/10.1016/j.jval.2010.10.015

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225–1231. https://doi.org/10.1016/s0895-4356(96)00002-9

Varian, H. R. (2014). Big data: New tricks for econometrics. The *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Vasquez, M. M., Hu, C., Roe, D. J., Chen, Z., Halonen, M., & Guerra, S. (2016). Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: Simulation and application. *BMC Medical Research Methodology*, *16*(1), 1–19. https://doi.org/10.1186/s12874-016-0254-8

Venables, W. N., & Ripley, B. D. (2002). Tree-based methods. In *Modern Applied Statistics with S. Statistics and Computing*. Springer.

Wailoo, A. J., Hernandez-Alava, M., Manca, A., Mejia, A., Ray, J., Crawford, B., Botteman, M., & Busschbach, J. (2017). 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*, *20*(1), 18–27. https://doi.org/10.1016/j.jval.2016.11.006

Wang, X., Yu, K., Dong, C., Tang, X., & Loy, C. C. (2019). Deep network interpolation for continuous imagery effect transition. In *Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition*.

Wu, E. Q., Mulani, P., Farrell, M. H., & Sleep, D. (2007). Mapping FACT-P and EORTC QLQ-C30 to patient health status measured by EQ-5D in metastatic hormone-refractory prostate cancer patients. *Value in Health*, *10*(5), 408–414. https://doi.org/10.1111/j.1524-4733.2007.00195.x

Yang, F., Devlin, N., & Luo, N. (2019). Impact of mapped EQ-5D utilities on cost-effectiveness analysis: In the case of dialysis treatments. *The European Journal of Health Economics*, *20*(1), 99–105. https://doi.org/10.1007/s10198-018-0987-x

Yang, F., Wong, C., Luo, N., Piercy, J., Moon, R., & Jackson, J. (2019). Mapping the kidney disease quality of life 36-item short form survey (KDQOL-36) to the EQ-5D-3L and the EQ-5D-5L in patients undergoing dialysis. *The European Journal of Health Economics*, *20*(8), 1195–1206. https://doi.org/10.1007/s10198-019-01088-5

Young, T. A., Mukuria, C., Rowen, D., Brazier, J. E., & Longworth, L. (2015). Mapping functions in health-related quality of life: Mapping from two cancer-specific health-related quality-of-life instruments to EQ-5D-3L. *Medical Decision Making*, *35*(7), 912–926. https://doi.org/10.1177/0272989x15587497

Yu, L., Wang, S., & Lai, K. K. (2007). A hybrid econometric-AI ensemble learning model for Chinese foreign trade prediction. In *Paper presented at the International Conference on Computational Science*.

Zar, J. H. (1972). Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, *67*(339), 578–580. https://doi.org/10.1080/01621459.1972.10481251

Zhao, S., Witten, D., & Shojaie, A. (2017). *In defense of the indefensible: A very naive approach to high-dimensional inference*. arXiv preprint arXiv:1705.05543.

Zheng, E., Tan, Y., Goes, P., Chellappa, R., Wu, D., Shaw, M., Sheng, O., & Gupta, A. (2017). When econometrics meets machine learning. *Data and Information Management*, *1*(2), 75–83. https://doi.org/10.1515/dim-2017-0012

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Aghdaee, M., Parkinson, B., Sinha, K., Gu, Y., Sharma, R., Olin, E., & Cutler, H. (2022). An examination of machine learning to map non-preference based patient reported outcome measures to health state utility values. *Health Economics*, *31*(8), 1525–1557. https://doi.org/10.1002/hec.4503

## APPENDIX A

### Index: tables included in the Appendix A

The mapping techniques used in this paper comply with the ISPOR Good Practices for Outcomes Research Task Force Report, and the Mapping onto Preference-Based Measures Reporting Standards (MAPS). Tables A1 and A2 provide the details of the two checklists.

**TABLE A1** Mapping onto preference-based measures reporting Standards (MAPS) checklist

| Section/topic | Item no. | Recommendation | Reported on page no. |
|---|---|---|---|
| Title and abstract | | | |
| Title | 1 | Identify the report as a study mapping between outcome measures. State the source measure(s) and generic, preference-based target measure(s) used in the study. | 1 |
| Abstract | 2 | Provide a structured abstract including, as applicable: Objectives; methods, including data sources and their key characteristics, outcome measures used and estimation and validation strategies; results, including indicators of model performance; conclusions; and implications of key findings. | 1 |
| Introduction | | | |
| Study rationale | 3 | Describe the rationale for the mapping study in the context of the broader evidence base. | 2–4 |
| Study objective | 4 | Specify the research question with reference to the source and target measures used and the disease or population context of the study. | 3–4 |
| Methods | | | |
| Estimation sample | 5 | Describe how the estimation sample was identified, why it was selected, the methods of recruitment and data collection, and its location(s) or setting(s). | 4 |
| External validation sample | 6 | If an external validation sample was used, the rationale for selection, the methods of recruitment and data collection, and its location(s) or setting(s) should be described. | NA |
| Source and target measures | 7 | Describe the source and target measures and the methods by which they were applied in the mapping study. | 4 |

(Continues)

TABLE A1 (Continued)

| Section/topic | Item no. | Recommendation | Reported on page no. |
|---|---|---|---|
| Exploratory data analysis | 8 | Describe the methods used to assess the degree of conceptual overlap between the source and target measures. | 8 |
| Missing data | 9 | State how much data were missing and how missing data were managed in the sample(s) used for the analyses. | NA |
| Modeling approaches | 10 | Describe and justify the statistical model(s) used to develop the mapping algorithm. | 5–8 |
| Estimation of predicted scores or utilities | 11 | Describe how predicted scores or utilities are estimated for each model specification. | 5–8 |
| Validation methods | 12 | Describe and justify the methods used to validate the mapping algorithm. | 5–8 |
| Measures of model performance | 13 | State and justify the measure(s) of model performance that determine the choice of the preferred model(s) and describe how these measures were estimated and applied. | 4 |
| Results | | | |
| Final sample size(s) | 14 | State the size of the estimation sample and any validation sample(s) used in the analyses (including both number of individuals and number of observations). | 8 |
| Descriptive information | 15 | Describe the characteristics of individuals in the sample(s) (or refer back to previous publications giving such information). Provide summary scores for source and target measures, and summarize results of analyses used to assess overlap between the source and target measures. | 8–9 |
| Model selection | 16 | State which model(s) is(are) preferred and justify why this(these) model(s) was(were) chosen. | 9–15 |
| Model coefficients | 17 | Provide all model coefficients and standard errors for the selected model(s). Provide clear guidance on how a user can calculate utility scores based on the outputs of the selected model(s). | Appendix C |
| Uncertainty | 18 | Report information that enables users to estimate standard errors around mean utility predictions and individual-level variability. | Appendix C&E |
| Model performance and face validity | 19 | Present results of model performance, such as measures of prediction accuracy and fit statistics for the selected model(s) in a table or in the text. Provide an assessment of face validity of the selected model(s). | Tables 2 and 3 |
| Discussion | | | |
| Comparisons with previous studies | 20 | Report details of previously published studies developing mapping algorithms between the same source and target measures and describe differences between the algorithms, in terms of model performance, predictions and coefficients, if applicable. | 15–16 |
| Study limitations | 21 | Outline the potential limitations of the mapping algorithm. | 16–17 |
| Scope of applications | 22 | Outline the clinical and research settings in which the mapping algorithm could be used. | 15–17 |
| Other | | | |
| Additional information | 23 | Describe the source(s) of funding and non-monetary support for the study, and the role of the funder(s) in its design, conduct and report. Report any conflicts of interest surrounding the roles of authors and funders. | 17 |

Abbreviation: NA, not applicable.

**TABLE A2** Mapping to estimate health-state utility from non-preference-based outcome measures: An ISPOR good practices for outcomes research task force report

| Recommendation | Reported |
|---|---|
| 1. Describe relevant differences between data sets that are candidates for mapping estimation. | One only dataset was used, which was collected for the purpose of this mapping study. |
| 2. Give full details of the selected data set. Describe how the study was run and patients were sampled. Provide baseline and follow up characteristics including the distribution of patients' disease severity. Missingness in the longitudinal pattern of responses should be described. | How the study was conducted and patients sampled provided in Section 2 (Data), patient characteristics provided in Table 1.<br>Data was cross-sectional with all questions mandatory, except for the Charlson comorbidity index (CCI), which was not used in the mapping study. Hence there was no missing data. |
| 3. Plot the distribution of the utility data. | Distribution of the observed versus predicted utilities presented in Figures 1–3. |
| 4. Justify the type of model(s) selected with reference to the characteristics of the target utility distribution and the proposed use of the mapping function. | Justification of models selected presented in Sections 3.3 and 3.4. |
| 5. Compare the dimensions of health covered by the target utility instrument and those covered by the explanatory clinical measure(s). | Description of instrument dimensions provided in Section 2. Spearman's rank correlation coefficients presented in Section 4.1. |
| 6. Describe the approach to determining the final model. Include tests conducted and judgments made. | Described in Section 3.2 |
| 7. Summary measures of fit are of limited value for the total sample. Provide information on fit conditional on disease severity as measured by the clinical outcome measure(s). A plot of mean predicted versus mean observed utility conditional on the clinical variable(s) should be included. | A range of summary measures are presented in Table 2. Distribution of the observed versus predicted utilities presented in Figures 1–3. |
| 8. Coefficient values, error term(s) distributions(s), variances, and covariances are required. | Presented in Appendix C and E. |
| 9. Provide an example predicted value for some sets of covariates. Consider providing a program that calculates predictions for user-defined inputs. | Examples of machine learning presented in Appendix B. Example of how to estimate predicted utility value presented in Appendix C. |
| 10. Parameter uncertainty in a mapping regression should be reflected using standard methods for Probabilistic Sensitivity Analysis (PSA). Assessment of model suitability for use in cost-effectiveness analysis should also consider the distribution of utility values for PSA, with particular focus on whether these lie outside the feasible utility range for the preference based measure (PBM). | Table 2 presents the proportion of observations truncated at one. |
| 11. When imputing data from a mapping function, individual-level variability should be incorporated using simulation methods and information about the distribution of the error term(s). These simulated data can be compared with the raw observed data, including an assessment of the range of values compared with the feasible range for the PBM. | Not applicable – no imputation conducted. |
| 12. Re-estimation of mapping results in a separate data set or other forms of validation are not routinely required. | Due to the lack of data on the five-level version of EQ-5D-5L, no external dataset was available, and only internal cross validation was applied in this study (mentioned in Section 3.2). |

*Note*: Summary of reporting of mapping studies recommendations.

## APPENDIX B

### Index: figures included in the Appendix

Figures B1–B5 illustrate examples of classification trees developed for different dimensions of EQ-5D-5L. It should be noted that a single tree does not produce the best algorithm for mapping. The classification tree for the anxiety and depression dimension of the EQ-5D-5L is also explained here to provide a better understanding of CART prediction algorithm:

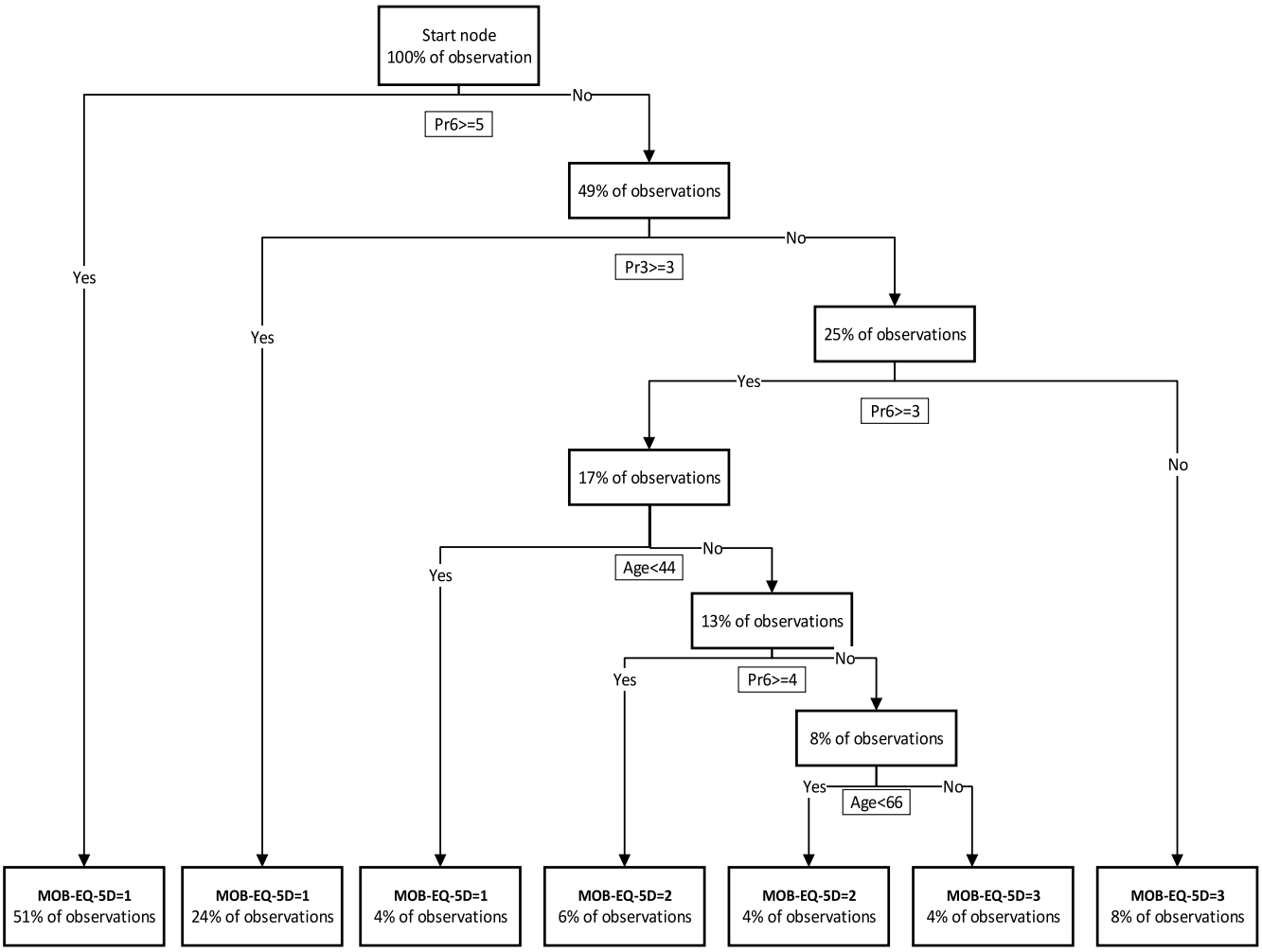


**FIGURE B1** Classification tree for Mobility dimension of EQ-5D-5L. Pr6: PROMIS-GH10, question 6: To what extent are you able to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair? 1: Not at all, 2: A little, 3: Moderately, 4: Mostly, 5: Completely. Pr3: PROMIS-GH10, question 3: In general, how would you rate your physical health? 1: Poor, 2: Fair, 3: Good, 4: Very good, 5: Excellent. MOB-EQ-5D: 1: I have no problems in walking about, 2: I have slight problems in walking about, 3: I have moderate problems in walking about,4: I have severe problems in walking about, 5: I am unable to walk about. PROMIS-GH10, PROMIS short form Global Health 10

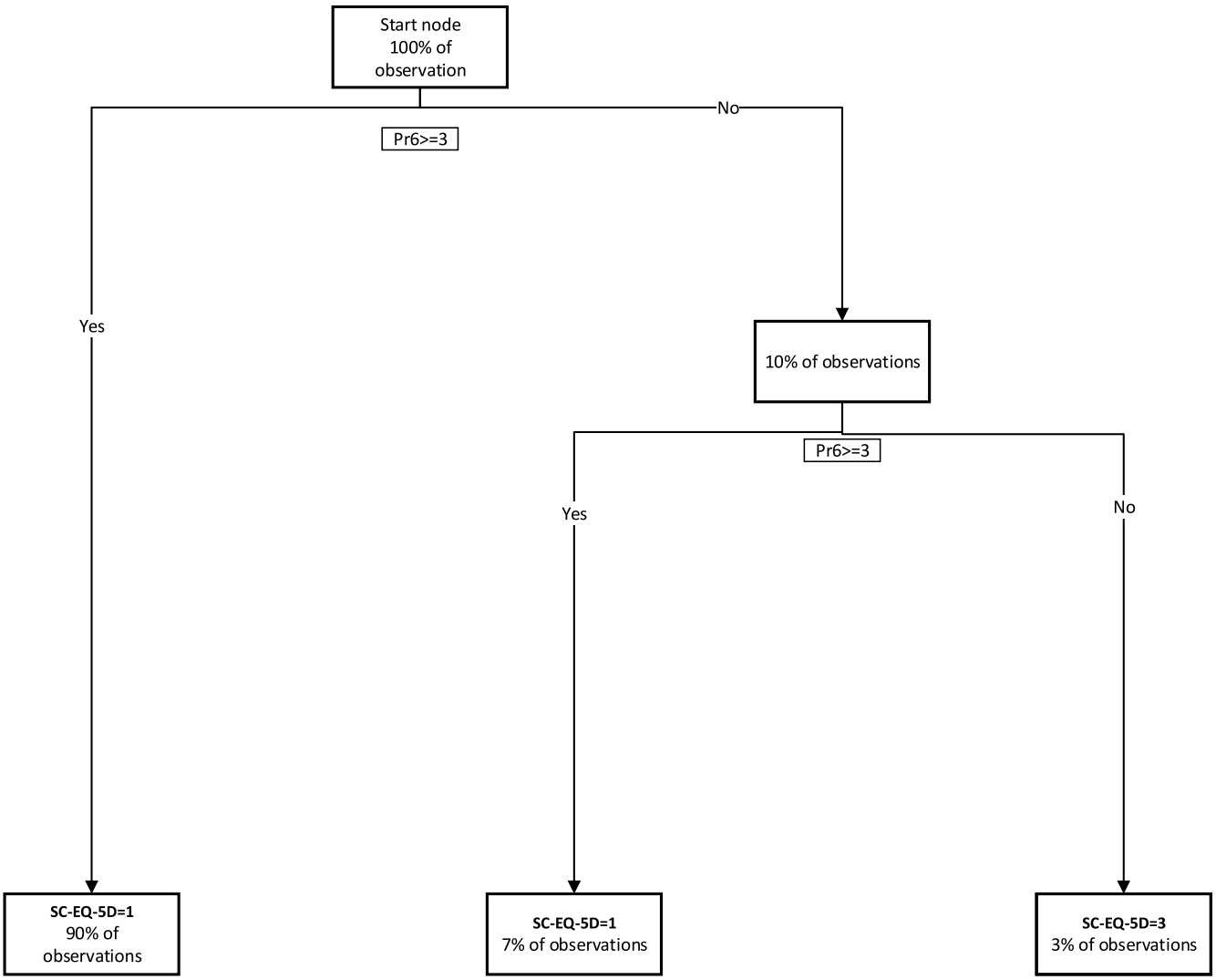


**FIGURE B2** Classification tree for Self-care dimension of EQ-5D-5L. Pr6: PROMIS-GH10, question 6: To what extent are you able to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair? 1: Not at all, 2: A little, 3: Moderately, 4: Mostly, 5: Completely SC-EQ-5D: 1: I have no problems washing or dressing myself, 2: I have slight problems washing or dressing myself, 3: I have moderate problems washing or dressing myself, 4: I have severe problems washing or dressing myself, 5: I am unable to wash or dress myself. PROMIS-GH10, PROMIS short form Global Health 10

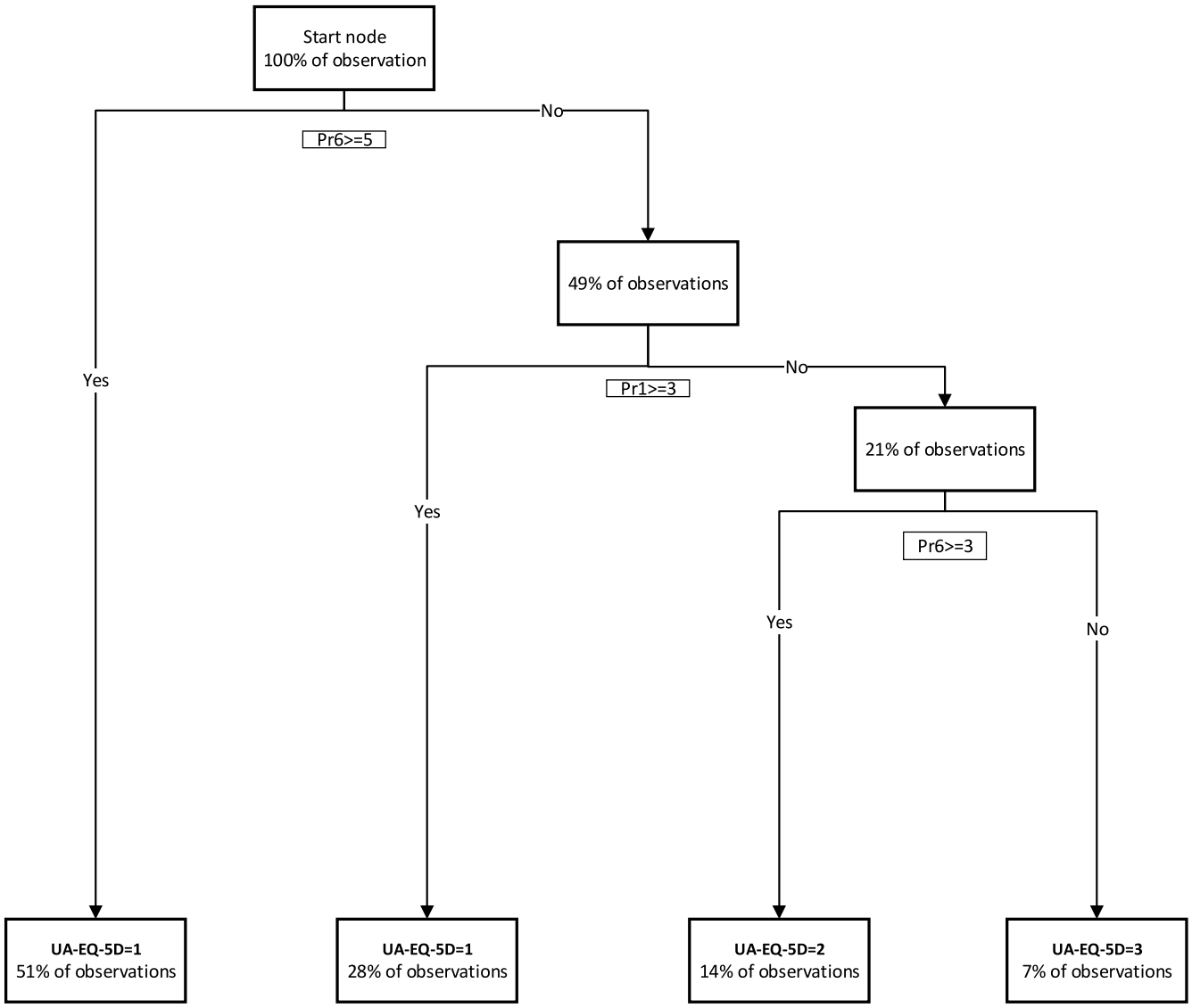


**FIGURE B3** Classification tree for usual activities dimension of EQ-5D-5L. Pr6: PROMIS-GH10, question 6: To what extent are you able to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair? 1: Not at all, 2: A little, 3: Moderately, 4: Mostly, 5: Completely. Pr1: PROMIS-GH10, question 1: In general, would you say your health is? 1: Poor, 2: Fair, 3: Good, 4: Very good, 5: Excellent. UA-EQ-5D: 1: I have no problems doing my usual activities, 2 I have slight problems doing my usual activities, 3: I have moderate problems doing my usual activities, 4: I have severe problems doing my usual activities, 5: I am unable to do my usual activities. PROMIS-GH10, PROMIS short form Global Health 10

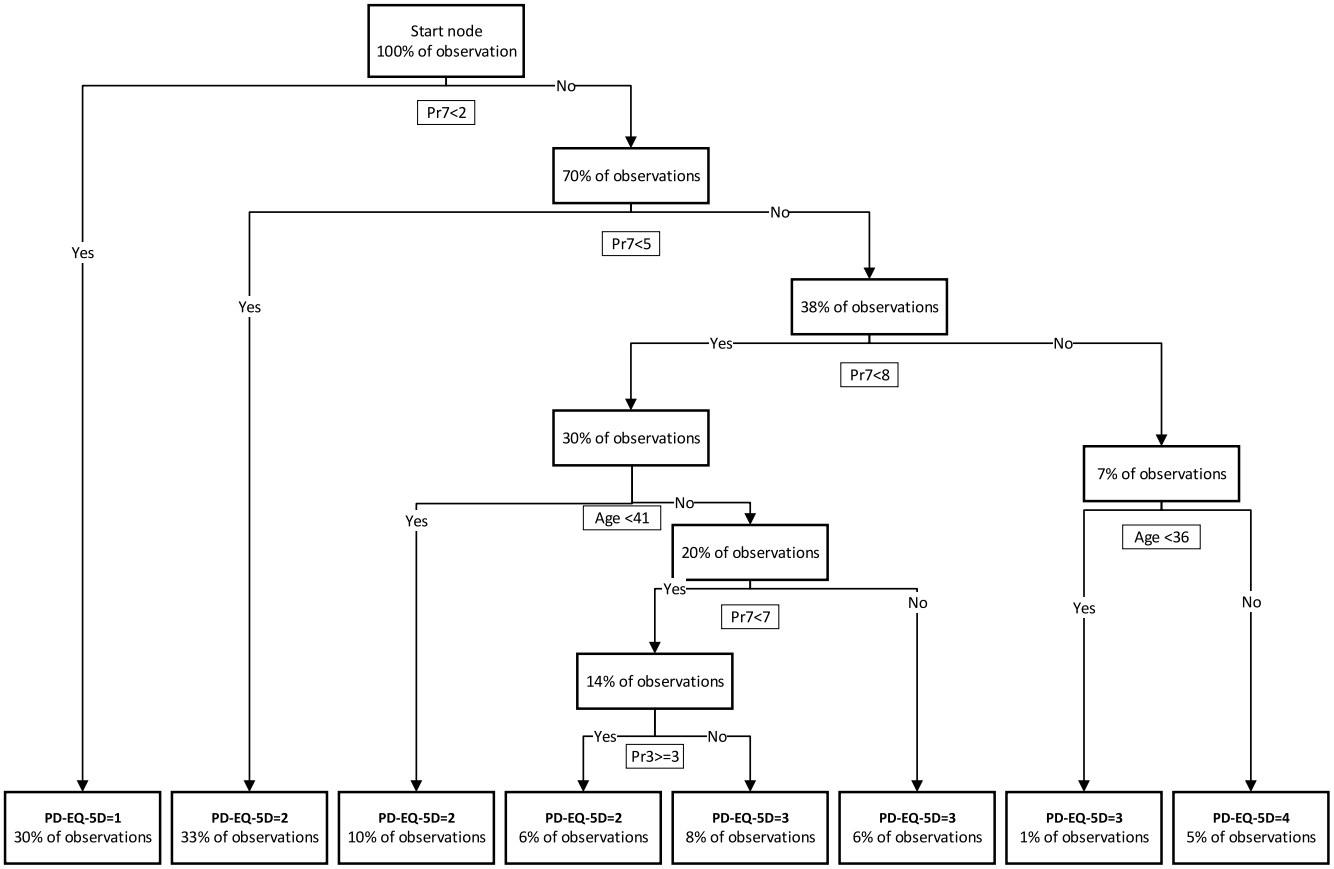


**FIGURE B4** Classification tree for Pain and Discomfort dimension of EQ-5D-5L. Pr7: PROMIS-GH10, question 7: In the past 7 days, how would you rate your pain on average? From 0 (no pain) to 10 (worst pain imaginable). Pr3: PROMIS-GH10, question 3: In general, how would you rate your physical health? 1: Poor, 2: Fair, 3: Good, 4: Very good, 5: Excellent. PD-EQ-5D: 1: I have no pain or discomfort, 2 I have slight pain or discomfort, 3: I have moderate pain or discomfort, 4: I have severe pain or discomfort, 5: I have extreme pain or discomfort. PROMIS-GH10, PROMIS short form Global Health 10

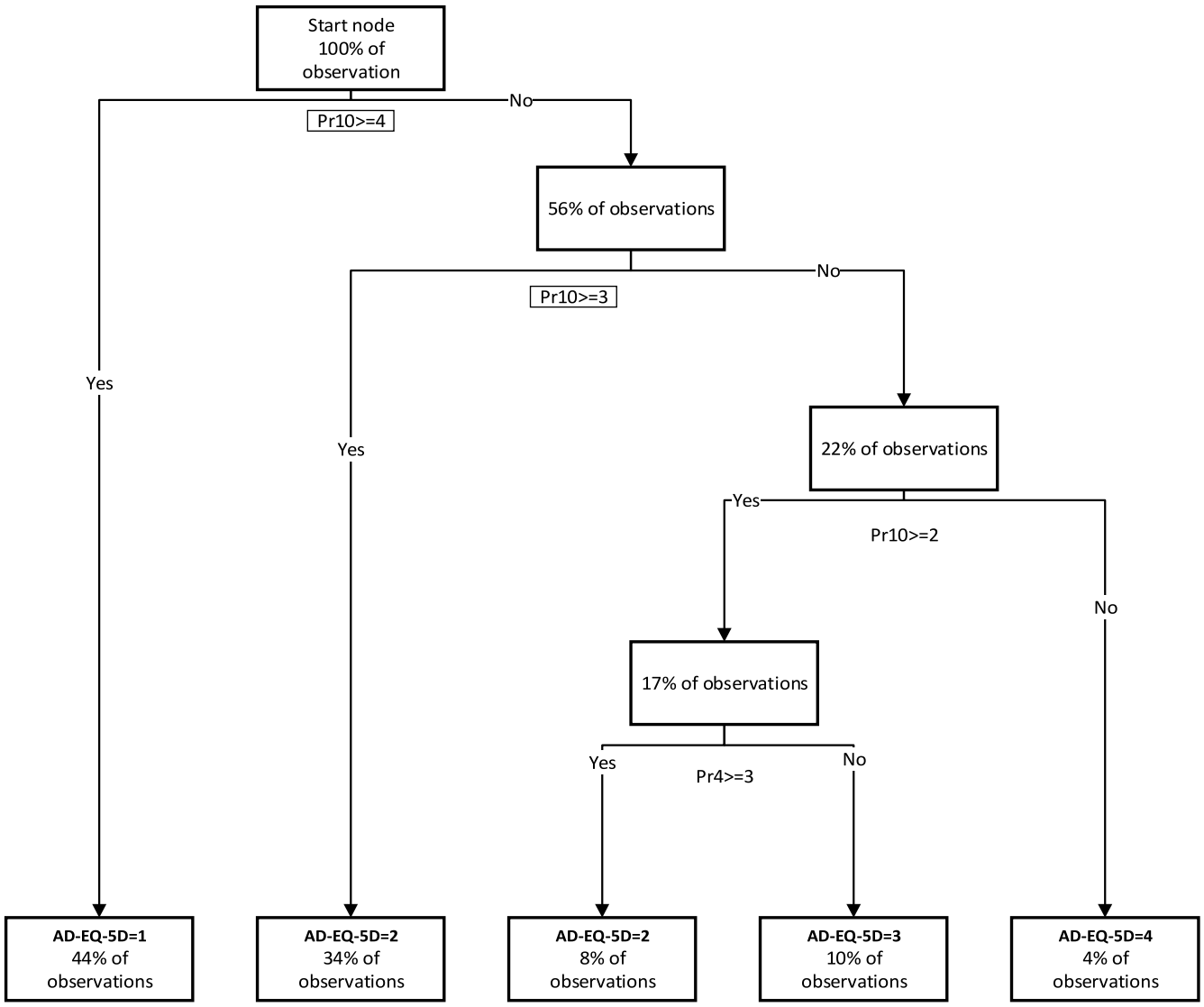


**FIGURE B5** Classification tree for Depression and Anxiety dimension of EQ-5D-5L. Pr10: PROMIS-GH10: In past 7 days, how often have you been bothered by emotional problems such as feeling anxious, depressed, or irritable? 1: Always, 2: Often, 3: Sometimes, 4: Rarely, 5: Never. Pr4: In general, how would you rate your mental health, including your mood and your ability to think? 1: Poor, 2: Fair, 3: Good, 4: Very good, 5: Excellent. AD-EQ-5D: 1: I am not anxious or depressed, 2: I am slightly anxious or depressed, 3: I am moderately anxious or depressed, 4: I am severely anxious or depressed, 5: I am extremely anxious or depressed

The pruned classification tree for the anxiety and depression dimension of EQ-5D-5L had five terminal nodes. PROMIS-GH10 question 10 had the principal role in data splitting and growing the tree. This concords with the nature of these questions as PROMIS-GH10 question 10 and the anxiety and depression dimension of EQ-5D-5L both ask the same question about feeling anxious, depressed or irritable. The tree predicted that respondents who either never or rarely experienced emotional problems (PROMIS-GH10 question 10 $\geq$ 4) responded with a score of one to the anxiety and depression dimension of EQ-5D-5L, reflecting that they were not depressed or anxious. Similarly, the classification tree suggested that respondents who were sometimes bothered emotionally (PROMIS-GH10 question 10 $\geq$ 3 and PROMIS-GH10 question 10 $<$ 4) were slightly anxious or depressed based on the EQ-5D-5L (anxiety and depression dimension of the EQ-5D-5L = 2).

Respondents who were often troubled by anxiety and depression (PROMIS-GH10 question 10 $\geq$ 2 and PROMIS-GH10 question 10 $<$ 3) were further classified by the PROMIS-GH10 question 4, which asks about mental health, mood and thinking ability. Ones who reported good to excellent mental health (PROMIS-GH10 question 4 $\geq$ 3) were classified as slightly anxious or depressed in EQ-5D-5L questionnaire (the anxiety and depression dimension of the EQ-5D-5L = 2). Those with lower

mental health status (PROMIS-GH10 question 4 < 3) were predicted to respond with a score of three to anxiety and depression dimension of the EQ-5D-5L.

Finally, the tree algorithm predicted that respondents who always experienced emotional issues (PROMIS-GH10 question 10 < 2) were severely depressed or anxious based on EQ-5D-5L (anxiety and depression dimension of EQ-5D-5L = 4).

The classification tree was unable to predict the number of extremely depressed or anxious (anxiety and depression dimension of EQ-5D-5L = 5) as only five people in the sample were in this category and this low number in the training sample was not sufficient to train the tree properly.

As it is reflected in the Figures B1–B5, the performance of trees for each dimension of EQ-5D-5L varied. As an example, the tree for Pain and Discomfort dimension, could predict people who ranked their pain one, two, three, and four. In the contrast the tree for Self-care dimension, could only predict ones in class 1 and 3. The reason for this variation was from the differences in the number of observations in the training sample in each class. The fewer, the number of observations in certain class, the ability of tree to predict the correct class accurately was poorer.

Moreover, the Self-care tree generated two leaves for a given class (level = 1). This was due the fact that there were more observations in that class in the training sample. Also, there were several criteria (questions or age interactions) to classify the observations.

An example of machine learning codes for generating Regression trees in R programming software

```
library(rpart)
library(rpart.plot)

#spliting data to training and testing sample

set.seed(111)
split = sample.split (all_data$eq-5d, SplitRatio = 0.5)
train = subset(all_data, split==TRUE)
test = subset(all_data, split==FALSE)

# growing a regression tree
tree _utility = rpart (formula= eq-5d ~ pr01 +pr02 +pr03 +pr04 +pr05 +pr06+ pr07+
pr08+ pr09+ pr10+age+gender, train, method="anova", ,control=rpart.control (mini-
split=10 ,cp=0.001, xval=10)

#predicting the utilities
tree_pred= predict(tree_utility, test)
hist(tree_pred)
hist(test$ eq-5d)
mean((tree_pred - test$ eq-5d)^2))
```

## APPENDIX C

**Index: tables and figures included in the Appendix C**

Appendix C provides two algorithms to calculate the utilities. One is based on adjusted limited dependent variable mixture model (ALDVMM) with explanatory variable set 3. The other is based on the Hybrid 1 model, which includes variables selected by the LASSO technique and is re-estimated with ALDVMM.

**Calculation of utilities**

Utilities are calculated from PROMIS-GH10 by matching responses to question and level specific coefficients reported in Tables C1 and C2 using the following formula:

EQ-5D-5L Utility = (Component one expected EQ-5D-5L utility × Component one probability) + (Component two expected EQ-5D-5L utility × Component two probability)

Where:

Component ($n$) expected EQ-5D-5L utility = $\sum Q_{i,k(n)} + \beta_{1(n)} * (Age) + \beta_{2(n)} * (age\ squared) + \beta_{3(n)} * Female + Constant\ (n)$

Where:

$Q_{i,k(n)}$: Coefficient Q for question i and level k in component n.

$\beta_{1(n)}$: Age coefficient in component n.

$\beta_{2(n)}$: Age squared coefficient in component n.

$\beta_{3(n)}$: Female coefficient in component n.

And,

$$\text{Component one probability} = \left(\frac{\exp(\text{Componant one probability constant})}{\exp(\text{Componant one probability constant})+1}\right)$$

$$\text{Component two probability} = \left(\frac{1}{\exp(\text{Componant one probability constant})+1)}\right)$$

**TABLE C1** Coefficients and standard errors from the best performing econometric model (Adjusted limited dependent variable mixture model [ALVDMM])

| Predictor variables | Component one coefficients | Standard errors | Component two coefficients | Standard errors |
|---|---|---|---|---|
| PROMIS-GH10 Q1 | | | | |
| Level-1 | −0.385959 | 0.0207294 | 0.0925135 | 0.0794249 |
| Level-2 | −0.0009395 | 0.0111615 | 0.018756 | 0.0642693 |
| Level-3 | 0.0027596 | 0.0095323 | 0.0402084 | 0.0565958 |
| Level-4 | −0.00748 | 0.0082035 | 0.0325671 | 0.0520424 |
| PROMIS-GH10 Q2 | | | | |
| Level-1 | −0.0585196 | 0.0139972 | −0.0794948 | 0.0714324 |
| Level-2 | −0.0070622 | 0.0093945 | −0.0291088 | 0.0569722 |
| Level-3 | −0.0066033 | 0.0082136 | −0.0320562 | 0.0509567 |
| Level-4 | −0.0058776 | 0.0072301 | −0.0638011 | 0.0457145 |
| PROMIS-GH10 Q3 | | | | |
| Level-1 | −0.0493412 | 0.0190513 | −0.0260497 | 0.0764026 |
| Level-2 | −0.0097279 | 0.0104587 | 0.002096 | 0.0630417 |
| Level-3 | 0.0061863 | 0.0095075 | 0.0626543 | 0.0583958 |
| Level-4 | 0.009719 | 0.0082286 | 0.1288012 | 0.0550993 |
| PROMIS-GH10 Q4 | | | | |
| Level-1 | −0.0192356 | 0.0109768 | −0.1073274 | 0.0579867 |
| Level-2 | −0.0156022 | 0.0076756 | 0.0167262 | 0.0484853 |
| Level-3 | −0.0012259 | 0.0065005 | 0.0983137 | 0.0424271 |
| Level-4 | −0.0038548 | 0.0056822 | 0.0472543 | 0.0368971 |
| PROMIS-GH10 Q5 | | | | |
| Level-1 | −0.0055637 | 0.0098399 | −0.0332379 | 0.059465 |
| Level-2 | −0.0088215 | 0.0081152 | 0.0007317 | 0.0532768 |
| Level-3 | −0.0104855 | 0.0074152 | 0.0262982 | 0.0486364 |
| Level-4 | −0.0085666 | 0.0066697 | 0.0076209 | 0.0428807 |
| PROMIS-GH10 Q6 | | | | |
| Level-1 | 0.5729949 | 0.051003 | −0.4978277 | 0.0734452 |
| Level-2 | −0.0321034 | 0.0131572 | −0.2048257 | 0.0383186 |
| Level-3 | −0.0295483 | 0.0063002 | −0.0573564 | 0.029585 |
| Level-4 | −0.0185259 | 0.0045397 | −0.0534656 | 0.0284769 |
| PROMIS-GH10 Q7 | | | | |
| Level-1 | −0.0359491 | 0.0058377 | −0.0286145 | 0.048754 |
| Level-2 | −0.0489133 | 0.006043 | −0.1483418 | 0.0453403 |
| Level-3 | −0.0557849 | 0.0062876 | −0.2193954 | 0.0464523 |

TABLE C1 (Continued)

| Predictor variables | Component one coefficients | Standard errors | Component two coefficients | Standard errors |
|---|---|---|---|---|
| Level-4 | −0.0648232 | 0.0073438 | −0.2431705 | 0.0518035 |
| Level-5 | −0.0737154 | 0.0080798 | −0.2560148 | 0.0444581 |
| Level-6 | −0.0957117 | 0.0088171 | −0.2982205 | 0.046128 |
| Level-7 | −0.1336154 | 0.0096407 | −0.3547439 | 0.04672 |
| Level-8 | −0.583629 | 0.0197676 | −0.4905812 | 0.0504853 |
| Level-9 | −0.1927814 | 0.0229887 | −0.7341895 | 0.0704952 |
| Level-10 | −1.099365 | 0.0245806 | −0.4552999 | 0.0904497 |
| PROMIS-GH10 Q8 | | | | |
| Level-1 | −0.0309681 | 0.0143363 | −0.13651 | 0.075945 |
| Level-2 | −0.0336167 | 0.0089216 | −0.2188279 | 0.0544924 |
| Level-3 | −0.0107356 | 0.0065049 | −0.169407 | 0.0489464 |
| Level-4 | −0.0088674 | 0.0061089 | −0.1161397 | 0.0479181 |
| PROMIS-GH10 Q9 | | | | |
| Level-1 | −0.013529 | 0.0160572 | −0.0730294 | 0.0648042 |
| Level-2 | 0.0015722 | 0.0087684 | 0.01400000 | 0.0536675 |
| Level-3 | 0.0013122 | 0.0072035 | −0.0279388 | 0.0471441 |
| Level-4 | 0.0017452 | 0.0063561 | 0.0180966 | 0.0431105 |
| PROMIS-GH10 Q10 | | | | |
| Level-1 | −0.0165344 | 0.0130433 | −0.2807613 | 0.0587058 |
| Level-2 | −0.0495842 | 0.0073591 | −0.2297149 | 0.0439222 |
| Level-3 | −0.0330347 | 0.0059642 | −0.0321068 | 0.0374949 |
| Level-4 | −0.0139751 | 0.0054585 | 0.0172828 | 0.0368554 |
| Age | 0.0003956 | 0.0005682 | 0.0043447 | 0.0033483 |
| Age squared | −9.99E-06 | 5.71E-06 | −0.0000673 | 0.0000338 |
| Female | 0.0009506 | 0.0035428 | −0.0446998 | 0.0201271 |
| Constant | 0.9745076 | 0.0168771 | 0.97559610 | 0.0993999 |
| Probability -Component 1 | | | | |
| Constant | 0.1232367 | 0.0777262 | | |
| /lns_1 | −3.240571 | 0.0435271 | | |
| /lns_2 | −1.403876 | 0.0336056 | | |
| sigma1 | 0.0391416 | 0.0017037 | | |
| sigma2 | 0.2456429 | 0.008255 | | |

*Note*: PROMIS-GH10 Q $n$ = nth question of PROMIS-GH10. The algorithm is based on ALVDMM set (3) that included PROMIS-GH10 questions as items, age, age squared and sex (Female = 1) as explanatory variables. For PROMIS-GH10 Q1, Q2, Q3, Q4, Q5, Q6, Q8, Q9, Q10 reference levels are level 5 and for PROMIS-GH10 Q7 reference level is level 0.

TABLE C2 Coefficients and standard errors from the Hybrid 1

| Predictor variables | Component one coefficients | Standard errors | Component two coefficients | Standard errors |
|---|---|---|---|---|
| PROMIS-GH10 Q1 | | | | |
| Level-1 | −0.3433060 | 0.0148387 | 0.0514095 | 0.0570408 |
| Level-2 | −0.0293574 | 0.0090414 | 0.0184637 | 0.0470782 |
| Level-3 | −0.0093709 | 0.0075675 | 0.0806698 | 0.0432715 |
| Level-4 | −0.0137952 | 0.0073393 | 0.1219705 | 0.0431204 |
| PROMIS-GH10 Q4 | | | | |
| Level-1 | −0.0216585 | 0.0121445 | −0.0596429 | 0.0550486 |
| Level-2 | −0.0216042 | 0.0078642 | 0.0436002 | 0.0464942 |
| Level-3 | −0.0017283 | 0.0066908 | 0.1149595 | 0.0405404 |
| Level-4 | −0.0067162 | 0.0058649 | 0.0598039 | 0.034641 |
| PROMIS-GH10 Q5 | | | | |
| Level-1 | −0.0112022 | 0.0100856 | −1.18E-02 | 5.52E-02 |
| Level-2 | −0.0150583 | 0.0078866 | 0.0234542 | 0.0493677 |
| Level-3 | −0.0168738 | 0.0072386 | 0.0576217 | 0.0450475 |
| Level-4 | −0.0127415 | 0.006478 | 0.0196911 | 0.0400461 |
| PROMIS-GH10 Q6 | | | | |
| Level-1 | −0.8875440 | 0.0476422 | −0.2675420 | 0.0653272 |
| Level-2 | −0.0503005 | 0.0135963 | −0.2063312 | 0.0361192 |
| Level-3 | −0.032574 | 0.0069206 | −0.0737328 | 0.0281568 |
| Level-4 | −0.0193999 | 0.004604 | −0.0544085 | 0.0275175 |
| PROMIS-GH10 Q7 | | | | |
| Level-1 | −0.0373753 | 0.0060229 | −0.0454425 | 0.0457808 |
| Level-2 | −0.050612 | 0.0061428 | −0.1630828 | 0.0425605 |
| Level-3 | −0.0583187 | 0.0063182 | −0.2345283 | 0.0439714 |
| Level-4 | −0.0679572 | 0.0074092 | −0.2694051 | 0.0491968 |
| Level-5 | −0.0763415 | 0.0080467 | −0.2772628 | 0.0416786 |
| Level-6 | −0.0977318 | 0.0086558 | −0.3293773 | 0.0435022 |
| Level-7 | −0.1341808 | 0.0097872 | −0.3876107 | 0.0447105 |
| Level-8 | −0.1393537 | 0.0265474 | −0.6160806 | 0.0480539 |
| Level-9 | −0.2173441 | 0.0232659 | −0.7665818 | 0.0676771 |
| Level-10 | −0.0343796 | 0.0450647 | −0.6329282 | 0.0739894 |
| PROMIS-GH10 Q9 | | | | |
| Level-1 | −0.0179659 | 0.0155147 | −0.0987504 | 0.0611147 |
| Level-2 | −0.0016861 | 0.0093691 | −0.0312816 | 0.050512 |
| Level-3 | 0.0066662 | 0.0072761 | −0.0893566 | 0.0449948 |
| Level-4 | 0.0039562 | 0.0062173 | −0.0171913 | 0.0399247 |
| PROMIS-GH10 Q10 | | | | |
| Level-1 | −0.0350141 | 0.01417 | −0.4081824 | 0.0529913 |
| Level-2 | −0.0561646 | 0.0071169 | −0.3029859 | 0.0403542 |
| Level-3 | −0.0360541 | 0.0060193 | −0.0912567 | 0.0349469 |
| Level-4 | −0.0150028 | 0.0055922 | −0.0184979 | 0.0343078 |
| Age | 0.0004218 | 0.0005884 | 0.003671 | 0.0032549 |
| Age squared | −0.00001 | 5.89E-06 | −0.0000534 | 0.0000328 |
| Female | 0.0001425 | 0.0036659 | −0.0395839 | 0.0194046 |
| Constant | 0.9989450 | 0.0167456 | 0.98656421 | 0.0882295 |

TABLE C2 (Continued)

| Predictor variables | Component one coefficients | Standard errors | Component two coefficients | Standard errors |
|---|---|---|---|---|
| Probability -Component 1 | | | | |
| Constant | 0.0898711 | 0.0818972 | | |
| /lns_1 | −3.217258 | 0.0491279 | | |
| /lns_2 | −1.423268 | 0.0335654 | | |
| sigma1 | 0.0400648 | 0.0019683 | | |
| sigma2 | 0.2409255 | 0.0080868 | | |

*Note*: PROMIS-GH10 Q $n$ = nth question of PROMIS-GH10. The algorithm is based on ALVDMM set (3) that included PROMIS-GH10 questions as items, age, age squared and sex (Female = 1) as explanatory variables. For PROMIS-GH10 Q1, Q4, Q5, Q6, Q9, Q10 reference levels are level 5 and for PROMIS-GH10 Q7 reference level is level 0.

## APPENDIX D

### Index: tables and figures included in the Appendix D

Table D1 provides the details of the performance of indirect approaches in each dimension of EQ-5D-5L.

TABLE D1 Goodness of fit for indirect approaches

| | Mobility | Self-care | Usual activity | Pain and discomfort | Anxiety and depression |
|---|---|---|---|---|---|
| Indirect mapping approaches | | | | | |
| Glogit | 66.35% | 75.96% | 71.83% | 85.30% | 54.87% |
| CART (classification trees) | 61.70% | 72.15% | 68.54% | 82.15% | 53.52% |
| Random forests | 66.67% | 75.92% | 72.53% | 85.92% | 55.87% |
| Bagging | 63.91% | 73.49% | 72.49% | 83.35% | 55.87% |
| NN | 68.08% | 76.53% | 73.24% | 86.35% | 57.75% |
| LASSO 1 | 69.48% | 76.85% | 73.24% | 86.38% | 58.22% |

*Note*: The table presents the percentage of correctly predicted for each dimension of EQ-5D-5L. LASSO 1, LASSO technique is used for prediction. Explanatory variables (without interactions) are only considered.

Abbreviations: GLOGIT, generalized logistic regression; LASSO, least absolute shrinkage and selection operator; NN, neural networks.