Mapping of the health of the Nation Outcomes Scale (HoNOS) and Positive and Negative Symptom Scale (PANSS) to the EQ-5D-3L in psychosis patients using the PHAMOUS study

Taylor & Francis
Taylor & Francis Group

Expert Review of Pharmacoeconomics & Outcomes Research

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ierp20

Mapping of the health of the Nation Outcomes Scale (HoNOS) and Positive and Negative Symptom Scale (PANSS) to the EQ-5D-3L in psychosis patients using the PHAMOUS study

Anne Kleijburg, Ben Wijnen, Wouter den Hollander, Silvia Evers, PHAMOUS investigators, Hans Kroon & Joran Lokkerbol

To cite this article: Anne Kleijburg, Ben Wijnen, Wouter den Hollander, Silvia Evers, PHAMOUS investigators, Hans Kroon & Joran Lokkerbol (06 Nov 2024): Mapping of the health of the Nation Outcomes Scale (HoNOS) and Positive and Negative Symptom Scale (PANSS) to the EQ-5D-3L in psychosis patients using the PHAMOUS study, Expert Review of Pharmacoeconomics & Outcomes Research, DOI: 10.1080/14737167.2024.2422456

To link to this article: https://doi.org/10.1080/14737167.2024.2422456

EXPERT REVIEW OF PHARMACOECONOMICS & OUTCOMES RESEARCH
https://doi.org/10.1080/14737167.2024.2422456

Taylor & Francis
Taylor & Francis Group

ORIGINAL RESEARCH

OPEN ACCESS Check for updates

Mapping of the health of the Nation Outcomes Scale (HoNOS) and Positive and Negative Symptom Scale (PANSS) to the EQ-5D-3L in psychosis patients using the PHAMOUS study

Anne Kleijburg [a,b], Ben Wijnen[b], Wouter den Hollander[c], Silvia Evers[a,b], PHAMOUS investigators*, Hans Kroon[d,e] and Joran Lokkerbol[b]

[a]Department of Health Services Research, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands; [b]Centre of Economic Evaluations, Trimbos Institute, Netherlands Institute of Mental Health and Addiction, Utrecht, The Netherlands; [c]Department of Epidemiology, Trimbos Institute, Netherlands Institute of Mental Health and Addiction, Utrecht, The Netherlands; [d]Department of Social and Behavioural Sciences, Tranzo Scientific Center for Care and Welfare, Tilburg University, Tilburg, Netherlands; [e]Department of Reintegration and Community Care, Trimbos Institute, Netherlands Institute of Mental Health and Addiction, Utrecht, The Netherlands

ABSTRACT

Objectives: When health outcomes relevant for economic evaluations are unavailable, algorithms can be developed to map utilities using available clinical outcome measures. This study aims to develop two mapping algorithms estimating EuroQol-5 dimension-3 level (EQ-5D-3 L) utilities using the clinician-rated Health of the Nation Outcome Scores (HoNOS) and Positive and Negative Syndrome Scale (PANNS).

Methods: A dataset with 2,029 observations of patients with psychotic disorders included EQ-5D-3 L, HoNOS, PANSS item scores, and demographics. Correlations between instruments were evaluated. Least Absolute Shrinkage and Selection Operator (LASSO) regression and random forest (RF) algorithms with various predictor variable sets were applied. Model performance was cross-validated using R-squared and Root Mean Square Error (RMSE).

Results: High ceiling effects were observed for EQ-5D-3 L, with weak to moderate negative correlations between EQ-5D and HoNOS ($r = -0.34$) and PANSS ($r = -0.27$). Overall, LASSO models outperformed RF models, with individual item models performing best for the HoNOS and PANSS, with the best observed RMSEs of 0.241 and 0.231, respectively.

Conclusions: The HoNOS and PANSS could be mapped onto EQ-5D-3 L utilities but lack accuracy for individual patient predictions. However, in the absence of alternatives, they could adequately predict population-based utility score differences for health economic evaluations.

ARTICLE HISTORY
Received 25 April 2024
Accepted 19 October 2024

KEYWORDS
HoNOS; PANSS; EQ-5D; psychotic disorders; severe mental illness; mapping; prediction

1. Introduction

Severe mental illness (SMI) is defined as the suffering from a psychiatric disorder for at least two years, which substantially interferes with multiple domains in life, thereby negatively impacting determinants of well-being such as educational attainment, work productivity, financial stability, housing, maintaining personal relations and life expectancy [1]. People diagnosed with disorders such as schizophrenia or other psychotic disorders, severe depression, or bipolar disorders commonly adhere to the criteria of a SMI. In the Netherlands, the prevalence of people suffering from SMI is estimated at 1.7% [2]. Due to its high disease burden, in addition to long term and intensive use of mental health services, an individual's need for social support and functioning are impacted, introducing high healthcare and societal costs resulting in a significant economic burden for both patients and society.

Across European countries, the total costs of mental illnesses have currently been estimated at being more than 4% of the Gross Domestic Product (GDP) [3]. When looking at healthcare expenditures, in some countries such as the Netherlands, the costs related to mental illness account for almost one-third of the total healthcare budget [4]. In addition to that, a recent German study has shown that costs associated with SMI are on average 2–3 times higher than that of non-severe mental illnesses [5].

In an attempt to limit healthcare expenditures, health economists try to guide the allocation of scarce resources by

CONTACT Anne Kleijburg akleijburg@trimbos.nl Department of Health Services Research, CAPHRI Care and Public Health Research Institute, Maastricht University, Da Costakade 45, P.O. Box 725, Utrecht, AS 3500, The Netherlands
*The PHAMOUS investigators involved were: S. Castelein[a], F. Jörg[b], G.H.M. Pijnenborg[c], W. Veling[d], E. Visser[b].
[a]Lentis Psychiatric Institute, Lentis Research, Groningen, the Netherlands; (2) University of Groningen, Faculty of Behavioural and Social Sciences, Department of Clinical Psychology and Experimental Psychopathology, Groningen, theNetherlands. [b]University of Groningen, University Medical Center Groningen, University Center for Psychiatry, Rob Giel Research center,Groningen, the Netherlands. [c]University of Groningen, Faculty of Behavioural and Social Sciences, Department of Clinical Psychology and ExperimentalPsychopathology, Groningen, the Netherlands; GGZ Drenthe Mental Health Institution, Assen, the Netherlands. [d]University of Groningen, University Medical Center Groningen, University Center for Psychiatry, Groningen, theNetherlands.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/14737167.2024.2422456

performing health economic evaluations, comparing treatment alternatives by looking at their incremental costs and (health) effects. To facilitate comparability of health effects between alternatives, national bodies for health economics often require the use of standardized health measures [6,7]. A commonly used measure is the Quality-Adjusted Life Year (QALY) which combines health-related quality of life (HRQoL) and survival in a single index. HRQoL is expressed using utility scores, where a score of 1 represents perfect health, a score of 0 represents dead, and a negative score represents a state worse than dead. Utilities can be measured using preference-based (multi-attribute) instruments such as the widely used EuroQol-5 Dimension (EQ-5D) or the Short Form-6 Dimension (SF-6D) [8]. These instruments are generic (i.e. not disease-specific), patient-reported, easy to use, and cover a wide range of health dimensions, supporting their generic application across diseases [9]. Many countries including the Netherlands require the use of the EQ-5D when performing evaluations according to the national health economic guidelines [7].

A frequently recognized limitation of these generic HRQoL instruments, however, is that depending on the disease, the dimensions covered in the instruments either lack relevance or sensitivity for capturing health-related changes. In the case of SMIs such as schizophrenia, the validity, and responsiveness of these generic patient-reported outcome measures (PROMs) is considered low [10,11]. Suggested reasons for this are the presence of ceiling effects (i.e. when individuals report to experience 'no problems' by lack of better representative response levels and dimensions, resulting in a full health score despite their condition), as well as the inability to capture the relevant domains of the patient's overall functioning (i.e. cognitive and clinical) and presentation of symptoms (i.e. positive, negative, and affective symptoms) [12,13]. Resulting from this lack in performance, and to minimize patient administrative burden, these generic HRQoL instruments are not included as outcome measures in routine clinical practice, and often also not during mental health studies.

Alternatively, a range of disease-specific instruments have are used which allow for improved validity and responsiveness when measuring functioning and/or symptoms of individuals with SMI in clinical studies and/or practice. These instruments consist of both patient-reported (e.g. the Schizophrenia Quality of Life Scale (SQLS)) and clinician-reported instruments (e.g. the Positive and Negative Syndrome Scale (PANSS)) [14,15]. Generally, clinician-reported measures are used to assess a patient's symptoms and functioning from the clinician's perspective, whereas PROMs often assess the impact on experienced HRQoL and wellbeing making them inherently more subjective [16]. As a result, clinician-reported measures are often considered more useful for evaluation of changes in a patient's health status, resulting in high implementation rates across both clinical practice and studies.

When performing an economic evaluation, however, the problem with clinician-reported instruments is that they cannot be used to derive utilities needed to perform QALY calculations for economic evaluations as often required by national guidelines [7,17]. This is because their valuation sets are not developed according to the preference-based principles used to construct generic utility measures [18]. In order for an instrument to classify as preference-based, each health state described by its classification system (*i.e.* each combination of items and levels) must be ordered by the preferences of respondents using a multi-attribute valuation study. In such valuation studies, a sample of respondents representative of the general population are presented with the possible health states and asked to indicate which health states are preferred over others. When combined and ordered by preferences, health states can each be assigned a utility score anchored between 0 and 1 (or even below 0), resulting in a (population/country-specific) valuation set reflective of how that specific population values various health states [19,20].

An increasingly applied solution to aid in the use of these non-preference-based measures in economic evaluations is the use of cross-walking or mapping techniques [21–24]. Here, mapping algorithms are used to predict the utility value generated by a generic preference-based PROM using the outcomes of a disease-specific measure when applied in the same patient population. To derive such a mapping algorithm various statistical methods (e.g. regression models) and machine learning techniques (e.g. random forests) can be applied to identify the best-performing model. However, ultimately model performance is largely determined by the level of conceptual overlap between mapped instruments [25,26].

In mental health services across Europe, the Health of the Nations Outcome Scales (HoNOS) is a clinician-rated instrument that is widely used in clinical practice for evaluating the health and social functioning of individuals with (severe) mental illness [27,28]. Given its broad applicability across mental health and psychiatric disorders, in a clinical context the HoNOS is generally considered a generic instrument; however, from a health economic perspective it would be considered disease-specific as it lacks domains sufficient to facilitate comparability across diseases areas. Similarly, the PANSS is disease-specific, clinician-reported instrument which more specifically aims to capture clinical symptoms in people with schizophrenia, which is also widely used in clinical practice and, in contrast to the HoNOS, also frequently used in clinical studies [15]. Ideally, to facilitate in the generation of health economic outcomes on care for individuals with SMI, the HoNOS and PANSS scores collected by mental health service providers or clinical studies can be used to map (EQ-5D-3 L) utilities.

For the HoNOS, no prior publications exist that report on its mapping to EQ-5D-3 L utilities; however, an earlier study in patients with Schizophrenia ($n = 304$) reported a moderate correlation ($r = -0.432$) between the instruments [29]. Furthermore, conceptual overlap appears to exist when comparing the EQ-5D domains and the items included in the HoNOS (i.e. depressed mood, daily activities, occupation, physical illness). For the PANSS, earlier studies have presented mapping algorithms of the PANSS to the EQ-5D (3 L and 5 L) in a European (mix of German, English and French patients) and Singaporean study population with mean scores indicating mild symptoms in people with schizophrenia [30,31]. However, since the use of such algorithms require that it was developed in highly similar populations or contexts, a mapping algorithm based on Dutch patients could provide an alternative for economic evaluations performed in the Dutch context.

Therefore, the primary aim of this study is to 1) explore the correlation between the instruments in individuals with SMI living in the Netherlands, and 2) develop an algorithm to map responses of the HoNOS to EQ-5D-3 L utilities for the use in economic evaluations. Furthermore, as a secondary aim, this study aims to develop an algorithm that maps PANSS responses available from our study sample to the EQ-5D-3 L.

2. Methods

2.1. Dataset

The current study is performed with data provided by the ongoing Pharmacotherapy Monitoring and Outcome Survey (PHAMOUS) cohort study [32]. Since 2006, PHAMOUS collects comprehensive Routine Outcome Monitoring (ROM) and screening data on people with (history of) psychotic disorders treated by one of four large mental healthcare organizations (Lentis, GGZ Friesland, GGZ Drenthe, University Center for Psychiatry of the University Medical Center Groningen) in the Netherlands. As described by Bartels-Velthuis *et al*. (2018), inclusion criteria for PHAMOUS were a diagnosis based on the Diagnostic and Statistical Manual 4th edition (DSM-4) or 5th edition (DSM-5) (primary or secondary) for schizophrenia, schizoaffective disorder, or other psychotic disorders at the time of inclusion, and a minimum age of 18 years. No exclusion criteria applied. PHAMOUS screenings occur on a yearly basis, are performed by trained nurses, and consists of social, psychiatric, and physical assessments. For the current analysis the HoNOS item scores, EQ-5D-3 L item scores, PANSS item scores and instrument assessment dates were selected, as well as age, sex, primary diagnoses (at the time of assessment) (DSM-IV or DSM-V), years since disease onset, EQ-VAS scores, living situation and daily activities as population descriptive variables. Selection criteria within the PHAMOUS database included all measurements that 1) adhere to the study's inclusion criteria, 2) include a completed HoNOS and EQ-5D assessment, which occurred no further than 90 days apart (to minimize capturing different health states over time), and 3) occurred between 2014 and 2024, excluding measurements older than 10 years (to minimize differences in health system and related changes over time). Due to the wide variation in possible primary diagnoses registered at the time of database extraction, DSM-4 and 5 primary diagnoses were recoded and categorized according to four major categories: 1) schizophrenic and psychotic disorders, 2) mood disorders, 3) anxiety, trauma, and stress-related disorders, and 4) other psychiatric disorders.

2.2. Instruments

The HoNOS is a clinician-reported, disease-specific instrument that can be used to assess the level of psychosocial functioning of patients with mental illness. The HoNOS is comprised of 12 items, which can be categorized into four subscales, being behavioral problems (items 1–3), cognitive and physical impairment (items 4–5), symptomatic problems (items 6–8), and social problems (items 9–12). Each item can be scored with a 0 (no problem), 1 (minor problem), 2 (mild problem), 3 (moderately severe problem), or 4 (severe to very severe problem). The final score is obtained by summing the scores of individual items totaling from 0 to 48, where lower scores indicate better outcomes [33]. For the HoNOS, total scores below 10 are considered interpretable as mild illness and above 13 as severe illness [34].

The PANSS is a clinician-reported, disease-specific instrument that can be used to assess severity of symptoms and psychopathology in patients with schizophrenia. The instrument consists of 30 items which can be categorized according to three domains, positive symptoms (7 items), negative symptoms (7 items), and general psychopathology (16 items). Each item can be scored using 7 levels, ranging from 1 (absent) to 7 (extreme). The total PANSS score and total domain scores are obtained by summing the scores of all items, where lower scores indicate better outcomes. This results in a range of a possible total PANSS score from 30 to 210, scores from 7 to 49 for the positive and negative domains, and a score from 16 to 112 for the general psychopathology domain [15,35]. For the PANSS, a total score of 58 is considered mild illness and a score of 116 as severe illness.

The EQ-5D-3 L is a patient-reported, generic HRQoL instrument that consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. In the 3 L version, each dimension is scored with one out of three responses, i.e. no problems (score of 1), moderate problems (score of 2), and severe problems (score of 3). As such, scores across the five domains may range from best possible (11111) to worst possible (33333) health, with a total of 243 possible health states [36,37]. When valued with the Dutch tariff for the EQ-5D-3 L, the best possible health is scored with a utility of 1, and the worst possible health is scored with a utility score of −0.3 representing a state that is worse than death (utility score of 0) [38]. In addition to scoring the five dimensions, the EQ-5D-3 L is accompanied by a visual analogue scale, the EQ-VAS, allowing patients to score their subjective health with endpoints ranging from 'best imaginable health state' (0) to 'worst imaginable health state' (100) [37].

2.3. Mapping analysis

To develop a mapping algorithm the following steps were taken:

2.3.1. Selecting predictors

To identify the best-performing mapping algorithm using the instrument items or subscales and demographics (age and sex), various model specifications were considered to predict the EQ-5D-3 L utility score. The EQ-VAS, diagnosis, and years since disease onset were not included in the primary analysis to prevent overcomplicating the algorithm and enhance its usability, and are only used as population descriptives. Based on the available combinations, the following models were considered: 1) all individual HoNOS/PANSS items, 2) all HoNOS/PANSS subscales, 3) all individual HoNOS/PANSS items and demographics, 4) all HoNOS/PANSS subscales and demographics, 5) all individual HoNOS/PANSS items, demographics and interactions, and 6) all HoNOS/PANSS subscales,

demographics, and interactions. Taking example of the HoNOS, this results in the following models:

(1) EQ-5D-3 L utility score = β0 + β1-12*HoNOS item 1-12
(2) EQ-5D-3 L utility score = β0 + β1-4*HoNOS subscale 1-4
(3) EQ-5D-3 L utility score = β0 + β1-12*HoNOS item 1-12 + β13*Age + β14*Sex
(4) EQ-5D-3 L utility score = β0 + β1-4*HoNOS subscale 1-4 + β5*Age + β6*Sex
(5) EQ-5D-3 L utility score = β0 + β1-12*HoNOS item 1-12 + β13*Age + β14*Sex + β15-26*HoNOS item 1-12*Age + β27-38*HoNOS item 1-12*Sex
(6) EQ-5D-3 L utility score = β0 + β1-4*HoNOS subscale 1-4 + β5*Age + β6*Sex + β7-10*HoNOS subscale 1-4*Age + β11-14*HoNOS subscale 1-4*Sex

2.3.2. Data preparation

To prepare the dataset for model fitting, first, the raw dataset was prepared by eliminating patient observations that were considered to have too many missing items in either measure. Given that the raw dataset is used to develop two separate mapping algorithms using either the HoNOS or the PANSS, here, elimination of missings results in two different sized datasets undergoing identical preparation steps. For the HoNOS dataset, no more than three items in each HONOS assessment were allowed to be missing for imputation, and no more than two items for EQ-5D-3 L. For the PANSS dataset, no more than two items from both the positive and negative domain, and four items from the general psychopathology domain in each PANSS assessment were allowed to be missing for imputation, as well as no more than two items for EQ-5D-3 L. Imputation of missing HoNOS, PANSS and EQ-5D-3 L items was performed using the within-patient row medians. Imputation of age was performed using the overall medians. Categorical variables with missing values (i.e. being NA) were imputed by replacing these with the label 'Missing' to prevent list-wise deletion from the analyses.

2.3.3. Data visualization

To improve our understanding of the data prior to analysis, data visualization of the HoNOS, PANSS and EQ-5D-3 L was performed. First, histogram plots were generated to visualize the distributions of scores across the study population. Next, total (and domain) HoNOS and PANSS scores against EQ-5D-3 L utility scores were plotted with a linear regression line, including 95% confidence intervals, and Spearman's correlation coefficients to investigate the correlation (also referred to as convergent validity) between the instruments, as well as a correlation matrix between individual HoNOS or PANSS items and EQ-5D items. Correlations coefficients can have either a positive or negative direction, and are considered weak if they are less than 0.3, moderate if r is between 0.3 and less than 0.7, and strong if r is 0.7 or higher [39].

2.3.4. Splitting the data in a train and test set

After imputation, the dataset was split to create a 'train' dataset, used to construct the mapping algorithms, and the 'test' dataset, used for the assessment of the models. The train set was randomly assigned 70% of the data and the test set 30%, using the function 'createDataPartition' from the R-package *caret* [40], which ensures that the outcome variable (*i.e.* EQ-5D utility) is equally distributed amongst both datasets.

2.3.5. Fitting of various mapping algorithms

Two mapping algorithms were considered: the Least Absolute Shrinkage and Selection Operator (LASSO) regression and the random forest (RF) algorithm. The LASSO algorithm was selected because its model structure is relatively easy to interpret, as well as its reported superiority for mapping utilities [41]. For RF, being a more flexible, 'black box' type of algorithm, model interpretation is more complex but mapping accuracy could be superior, especially on datasets with many variables and interaction terms.

The LASSO uses shrinkage and regularization to find the optimal balance between model accuracy and model parsimony/complexity. The LASSO is similar to linear regression in that it minimizes the sum of squares. However, by including a penalty term (the shrinkage parameter *lambda*) the predictor variables that are least important can be eliminated from the model [42]. To determine what value of lambda produces the optimal model, a grid of lambda values ($n = 200$) ranging from 0 to 1000 (where 0 equals no penalty and is identical to regular linear regression) were tested.

With RF a multitude of decision-trees are generated and aggregated which can then be used to predict utility. To generate the forest of individual trees, bootstrapping of the dataset is combined with random variable selection, which means that each split in each tree is limited to a random subset of predictors. As a result, each tree is well fitted to a subset of the predictors, which when aggregated to a forest, reduces the correlation between trees and reduces the variance. To find the best-performing random forest, tuning of the *mtry* parameter was applied, which specifies the size of the predictor variable subset and allows to find the optimal number of variables selected for optimizing predictive results. Model tuning for both LASSO and RF was performed using K-fold cross-validation (k = 10).

Model performance was assessed using the Root Mean Square Error (RMSE), representing the standard deviation of the prediction errors, and the R-squared, representing the amount of variance explained by the model relative to its total variance. Here, a lower RMSE and a higher R-squared represent better model performance. The RMSE was used as the leading performance metric in model assessment and selection. All analyses were performed in the statistical programming language R version 4.2.0, and using the *caret* package (version 6.0–9.3) for machine learning [40,43].

2.3.6. Evaluating the model on the test set

Based on performance measures, the best-performing models on the training set were selected to predict outcomes in the test set. Here, the RMSE was again used to select the best-performing model. The R-squared, the mean absolute error (MAE), and normalized mean absolute error (NMAE) are reported as additional performance indicators. The MAE is a linear measure indicating the difference between the original and predicted values, and the NMAE expresses the magnitude of this error relative to the original sample mean

facilitating interpretability and comparability between samples. For the best-performing LASSO models standard errors surrounding the regression coefficients were approximated using bootstrapping, as these cannot be computed directly due to regularization [44]. Additionally, to visualize the level of agreement between observed utilities and utility scores predicted by our mapping algorithm a Bland-Altman plot was constructed and assessed on the width of its limits, and the presence of trends and variability across plotted values.

2.4. Sensitivity analyses

To explore the impact of assumptions made during data preparation and variable selection for model specification on the model's performance, sensitivity analyses were performed by 1) using alternative imputation methods, and 2) including diagnosis as an additional predictive variable in the models along with the other included demographics (age and sex). Alternative imputation for predictor variables was performed using the k-Nearest Neighbors (kNN) method, available from the caret package. kNN imputes missing values with non-missing values of patients in the dataset with similar characteristics. As it is not possible to use kNN imputation on both the predictor and response (EQ-5D-3 L utilities) variables simultaneously (response variables are required for prediction), within-patient median imputation was used to impute the missing EQ-5D-3 L level scores, similar to the base-case analysis.

2.5. Ethical compliance

Research on this dataset has been deemed exempt from the Medical Research Involving Human Subjects Act (WMO) by the Medical Ethical Committee of the University Medical Center Groningen. This exemption was granted because no additional burden is placed on service users, the research involves preexisting routine healthcare data, and research on the data serves a broad public interest. Data collected by PHAMOUS is in accordance with the declaration of Helsinki. Data is collected using an integrated application in the electronic patient file system. This application aims to support shared decision making with the patient by composing shareable reports and treatment plans and is part of regular care. Patients can opt out from the use of their (anonymized) data for scientific research. PHAMOUS is part of regular care. Data is encrypted, pseudonymized, and linked by a trusted third party before being available for research purposes, ensuring secure and controlled access in line with the principles of privacy by design. Therefore, data collected by the PHAMOUS study does not contain directly identifiable patient information. Service users are informed that their pseudonymized data could be used for scientific research, and they have the option to opt out of data use.

3. Results

3.1. Patient characteristics

A total of 2,527 patient records were extracted from the PHAMOUS study. From these, 477 records were excluded for missing a primary diagnosis, causing uncertainty to their adherence to the selection criteria. From the HoNOS record sample 21 records were excluded for having too many missing values in the HoNOS or EQ-5D items, resulting in a final total sample of 2,029 included records. From the PANSS record sample 844 records were excluded for having too many missing values in the PANSS or EQ-5D items, and 21 for measurements being more than 90 days apart, resulting in a final total sample of 1164 included records. Excluded patient records had similar characteristics to the final sample, however, from the PANSS record sample, more records from patients with schizophrenic and psychotic disorders were included due to having complete PANSS measurements. For the HoNOS, the final record consisted of a total of 1,853 unique patients, where 128 patients had more than one measurement, equaling to 15% of records. For the PANSS, the final record consisted of a total of 1061 unique patients where 74 patients had more than one measurement, also equaling 15% of the records. Patient characteristics of both samples are presented in Table 1, and are based on the most recent record available for each patient. Overall, the samples consisted of more male patients, with a mean age of 50 years (SD ± 12 years), and the majority of patients had a primary diagnosis in the category of schizophrenic and psychotic disorders. Patients were reported to experience relatively little symptoms (mean PANSS total score = 48 out of 210 for both HoNOS and PANSS records) and a moderate below average HRQoL (mean EQ-5D utility of 0.72 and 0.75 for HoNOS and PANSS records respectively), when compared to the HRQoL of the Dutch general population (0.857 ± 0.183 at 50 years) [45]. For both the HoNOS and PANSS samples, the mean number of days between the measurements with both instruments was approximately 2 days, supporting that the expected impact of differences over time between measured health states is likely minimal.

3.2. Data visualization

Distributions of observed EQ-5D-3 L utility scores, total HoNOS scores, and total PANSS scores in the included data sample are presented in Figure 1. Here, for the EQ-5D (Figure 1(a)) it can be observed that for the majority of patient records (25%) a utility score of 1 for perfect health was reported, indicating the presence of ceiling effects. Visualization of the EQ-5D-3 L utility scores against total HoNOS (Figure 2) and EQ-5D-3 L utility scores against total PANSS scores (Supplementary Figure S1) with Spearman's r indicated a weak to moderate negative correlation of $r = -0.34$ ($p < 0.05$) and $r = -0.27$ ($p < 0.05$) between the measures, respectively. In addition to that, visualization of the self-rated EQ-5D and the clinician-rated HoNOS and PANSS clearly show large discrepancies between assessments. Patients frequently scored (close to) perfect health on the EQ-5D while in those instances the physicians' ratings ranged from agreeable ratings with scores to less agreeable ratings with higher scores, approaching total HoNOS scores as high as 31 and total PANSS scores as high as 93. Correlations matrices for the HoNOS and PANSS items, subscales, and total scores between individual EQ-5D items and utility scores are presented in Supplementary Figure S2 and S3. Overall, across both the HoNOS and PANSS, the majority of

Table 1. Patient characteristics.

	HoNOS (n=1,853)		PANSS (n=1,061)	
	Mean (SD)/%	Missing	Mean (SD)/%	Missing
Sex				
Male	61.1%		65.3%	
Female	38.5%		34.2%	
Missing	0.4%		0.5%	
Age	50.1 (12.4)	0.4%	50.4 (12.5)	0%
Category of primary diagnosis				
1. Schizophrenic and psychotic disorders	68.1%		82.4%	
2. Mood disorders	15.4%		9.6%	
3. Anxiety, trauma and stress-related disorders	7.1%		3.2%	
4. Other psychiatric disorders	9.4%		4.8%	
Years since disease onset	22.8 (12.7)	26.8%	22.9 (12.8)	15.2%
Living situation				
Alone	43.7%		48.8%	
With partner/children	17.6%		15.6%	
With family/others	5.4%		5.2%	
Supported housing	14.2%		16.5%	
Long-stay (clinical)	6.9%		5.1%	
Other	2.4%		2.3%	
Missing	9.8%		6.6%	
Daily activities				
Paid work	7.9%		8.9%	
Supported employment, voluntary work, education	23.2%		25.9%	
Household or other	24.4%		27.1%	
No activities	12.2%		12.8%	
Missing	32.3%		25.3%	
PANSS total	48.5 (14.9)	49.9%	48.2 (14.8)	13.0%
PANSS Positive	11.2 (4.47)	43.4%	11.1 (4.44)	2.4%
PANSS Negative	12.4 (5.28)	45.1%	12.4 (5.26)	5.1%
PANSS Psychopathology	24.9 (7.58)	47.0%	24.7 (7.48)	7.9%
HoNOS Total	8.71 (5.65)	3.6%	8.46 (5.53)	3.6%
HoNOS Behavior	1.01 (1.42)	0.8%	0.91 (1.31)	0.3%
HoNOS Functional	1.73 (1.56)	0.2%	1.68 (1.53)	0.3%
HoNOS Symptoms	3.18 (2.23)	1.3%	3.10 (2.30)	1.5%
HoNOS Social	2.87 (2.61)	1.6%	2.83 (2.54)	1.7%
EQ-5D-3L utility score	0.72 (0.27)	0.2%	0.75 (0.26)	0.2%
EQ-VAS	66.0 (17)	1.7%	67.9 (16.3)	1.8%
Days between HoNOS – EQ5D measurements	1.93 (27.8)	0%	2.27 (22.8)	0%
Days between PANSS – EQ5D measurements	2.25 (42.4)	41.1%	−0.56 (7.11)	0%

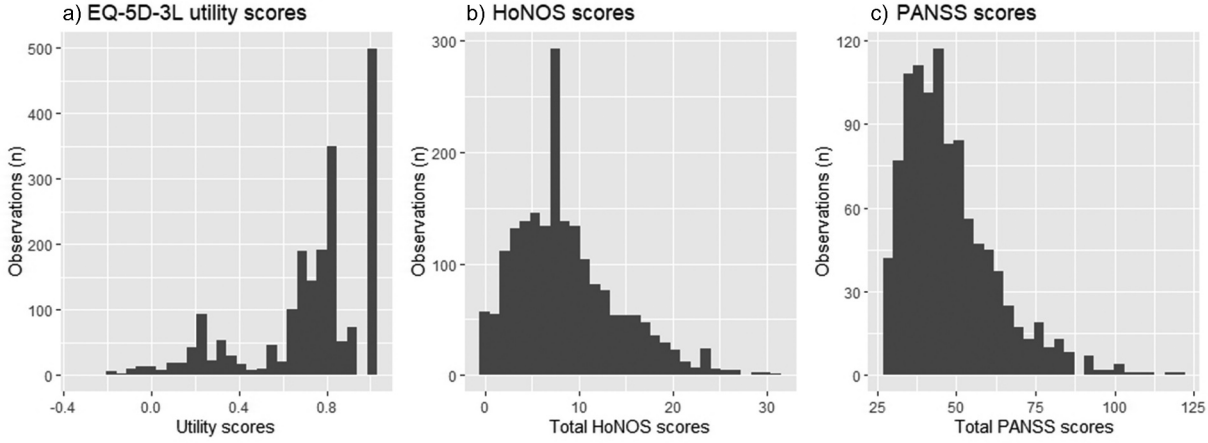

Figure 1. Histogram distribution plots of the EQ-5D-3L utility scores (a), and total HoNOS scores (b) for a total of 2,115 complete patient observations.

items show little correlation with the EQ-5D domains and utility scores. For the HoNOS, largest correlations are observed between HoNOS item 5 (physical illness) and EQ-5D item 1 (mobility) with $r = 0.42$, HoNOS item 7 (depressed mood) and EQ-5D item 5 (anxiety/depression) with $r = 0.41$, and HoNOS item 5 and EQ-5D item 4 (pain/discomfort) with $r = 0.41$. Similarly, largest correlations are observed between PANSS item G6 (Depression) and PANSS item G2 (Anxiety) with EQ-5D item 5 (anxiety/depression) with $r = 0.49$ and $r = 0.40$, respectively.

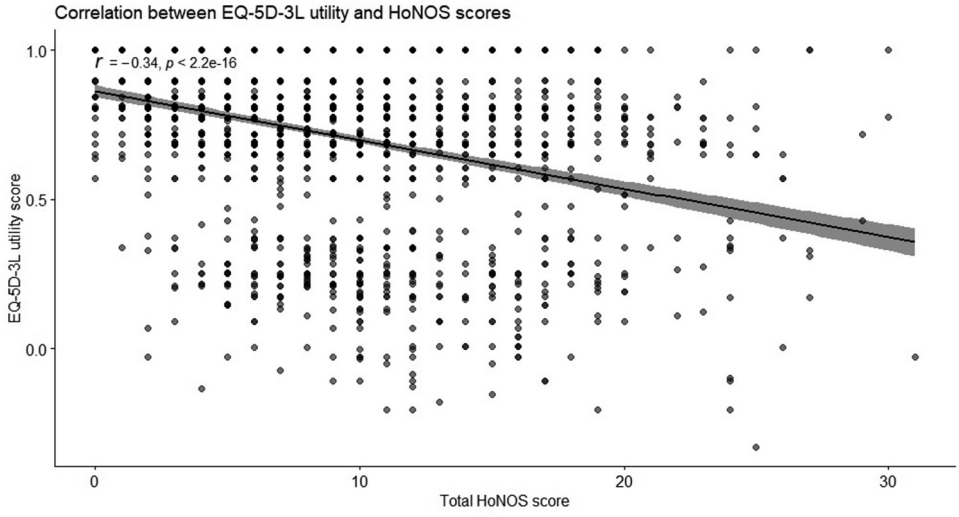

Figure 2. EQ-5D-3L utility scores are plotted against total HoNOS scores in patients with severe mental illness ($n = 2,115$) accompanied by a linear regression line, a 95% confidence interval and Spearman's correlation coefficient (r), indicating a negative and weak to moderate correlation ($r = -0.31$, $p < 0.05$).

3.3. Fitting of mapping algorithms and model evaluation

Model performance for each of the six predictor sets modeled with LASSO and RF are shown in Table 2. For both the LASSO and RF models, the predictor sets containing individual HoNOS/PANSS items showed superior performance to those with only subscales. Overall, for both the HoNOS and PANSS, best performance was seen for the LASSO models, with comparable R-squared and RMSE values for all three individual item models (Models 1, 3, and 5). The LASSO individual item models were selected for model prediction on the test set (shown in Table 3). Model performances on the test set were comparable to those of the train set, indicating the models did not overfit on the train data. Looking at the predicted EQ-5D utility scores, none of the models were able to predict full health, or scores that were close to 0 or negative. For the HoNOS, the largest range of predicted utilities was observed for Model 3 and ranged between a score of 0.248 and 0.907. For the PANSS, the largest range of predicted utilities was observed for Model 1 and ranged between a score of 0.253 and 0.912.

When looking at the RMSE, R-squared, and the MAE values found by the mapping algorithms on the test set, again all three LASSO models performed almost identically. Based on the performance measures and predicted ranges, whilst also opting for the simplest model (i.e. the model with the least number of included variables), Model 3 could be considered the optimal model for mapping both the HoNOS and PANSS. However, given the minimal differences in model performance and to facilitate use in absence of patient demographics, model coefficients for both Model 1 and Model 3 are presented (Table 4 and Supplementary Table S2). Coefficients were ordered according to the variable importance (VarImp) function from the caret package, providing the relative ranking of the contribution of each of the independent variables in determining the outcome. For the HoNOS, in both Model 1 and Model 3 the variables for HoNOS items 7 (depressed mood), item 5 (physical illnesses), and item 8 (mental/behavioral problems) are found most important. For the PANSS, in both Model 1 and Model 3 the variables for PANSS items G6 (depression), item G1 (somatic/physical illnesses), and item P3 (hallucinatory behavior) are found most important. For Model 3 of both the HoNOS and PANSS the demographic characteristic of being female is also included as a coefficient, slightly reducing their predicted HRQoL.

To explore the agreement between the observed utility scores in the dataset and the predicted utility scores using the coefficients found for Model 3, a Bland Altman plot was constructed for both the HoNOS (Figure 3) and PANSS algorithms (Supplementary Figure S4). For the HoNOS, a total of 93.6% of the measurements was found to be within the 95% limits of agreement, meaning that for

Table 2. Model performance on the training set.

	HoNOS				PANSS			
	LASSO		Random Forest		LASSO		Random Forest	
	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
Model 1: all items	0.216	0.235	0.184	0.239	0.254	0.218	0.206	0.225
Model 2: all subscales	0.154	0.244	0.102	0.258	0.118	0.239	0.092	0.248
Model 3: all items + demographics	0.213	0.235	0.187	0.239	0.261	0.217	0.208	0.225
Model 4: subscales + demographics	0.160	0.244	0.125	0.249	0.127	0.238	0.115	0.238
Model 5: all items + demographics + interactions	0.214	0.236	NA*	NA*	0.254	0.218	NA*	NA*
Model 6: subscales + demographics + interactions	0.155	0.244	NA*	NA*	0.130	0.238	NA*	NA*

Abbreviations: RMSE: Root Mean Square Error, LASSO: Least Absolute Shrinkage and Selection Operator, NA: Not applicable.
*Given the nature of the model/algorithm, interaction terms are automatically included when using a RF algorithm.

Table 3. Predicted vs observed EQ-5D-3 L utility scores.

	Observed EQ-5D-3L utility	Predicted EQ-5D-3L utility		
		Model 1: Items	Model 3: items +demographics	Model 5: Items + demographics + interactions
LASSO regression of the HoNOS				
Mean	0.720	0.723	0.723	0.723
Minimum	−0.329	0.276	0.248	0.256
Maximum	1	0.880	0.907	0.877
RMSE		0.242	0.241	0.241
MAE		0.187	0.187	0.186
NMAE		26.0%	26.0%	25.8%
R-Squared		0.254	0.261	0.263
LASSO regression of the PANSS				
Mean	0.751	0.757	0.757	0.757
Minimum	−0.204	0.253	0.258	0.249
Maximum	1	0.912	0.920	0.903
RMSE		0.232	0.231	0.231
MAE		0.174	0.172	0.172
NMAE		23.2%	22.9%	22.9%
R-Squared		0.186	0.192	0.193

Abbreviations: LASSO: Least Absolute Shrinkage and Selection Operator, RMSE: Root Mean Square Error, MAE: Mean Absolute Error.

Table 4. LASSO regression mean coefficients including standard errors (SEs) for models mapping the EQ-5D-3 L utility using only the individual HoNOS items (Model 1) and the individual HoNOS items and demographics (Model 3). Coefficients are ordered based on the variable importance identified by Model 3.

Model items included	Model 1 Coefficients	SE	Model 3 Coefficients	SE
(Intercept)	0.8775	0.0097	0.9004	0.0128
HoNOS item 7: Depressed mood	−0.0719	0.0078	−0.0747	0.0079
HoNOS item 5: Physical illness	−0.0517	0.0070	−0.0551	0.0072
HoNOS item 8: Mental/behavioural problems	−0.0260	0.0056	−0.0267	0.0057
Sex (Female)	n.a.	n.a.	−0.0223	0.0134
HoNOS item 4: Cognitive problems	–	–	0.0061	0.0070
HoNOS item 2: Self-harm	–	–	−0.0048	0.0103
HoNOS item 10: Activities of daily living	–	–	−0.0044	0.0052
HoNOS item 3: Alcohol/Substance abuse	–	–	−0.0041	0.0063
HoNOS item 6: Delusions or Hallucinations	–	–	−0.0036	0.0051
HoNOS item 12: Occupation	–	–	−0.0027	0.0050
HoNOS item 9: Relationship problems	−0.0005	0.0035	−0.0021	0.0046
Age	n.a.	n.a.	–	–
HoNOS item 11: Living conditions	–	–	–	–
Sex	n.a.	n.a.	–	–
HoNOS item 1: Overactive/aggressive behaviour	–		–	–

these measurements the difference between its observed and predicted utility score ranged between −0.47 and 0.46. Given the possible range of EQ-5D-3 L utility scores ranges between −0.3 and 1, the identified limits can be considered relatively wide, indicating that considerable variability can occur between utilities observed and predicted. Additionally, when looking at the plotted differences a trend with positive direction can be observed across increasing utility values (i.e. there is more overprediction for lower-end utilities, and more underprediction for higher-end utilities), and variability between differences decreases across increasing utility values. Overall, agreement between predicted and observed utility scores appears largest between a value of ~0.6 and ~0.8. Looking at the Bland-Altman plot constructed using Model 3 trained with the PANSS, observed conclusions are identical.

3.4. Sensitivity analyses

Model performances on the training set for models tested as part of additional sensitivity analyses are presented in Supplementary Table S3. The first sensitivity analysis showed that the inclusion of patient diagnosis as predictor variable (in addition to the demographics age and sex) introduced only minor improvements in predictive performance for the specified LASSO models of both the HoNOS and PANSS. When comparing the RF models with and without diagnosis, predictive performance showed larger improvements upon including diagnosis; however, overall performance remains poorer when compared to the LASSO models. When looking at the second set of sensitivity analyses using kNN imputation, again no meaningful improvements are observed in the LASSO and RF model performances for both the PANSS and HoNOS, indicating imputation methods have little impact on model performance using this dataset.

4. Discussion

The primary aim of this study was to develop a mapping algorithm for the EQ-5D-3 L using the HoNOS in patients with SMI, and explore the correlation between the

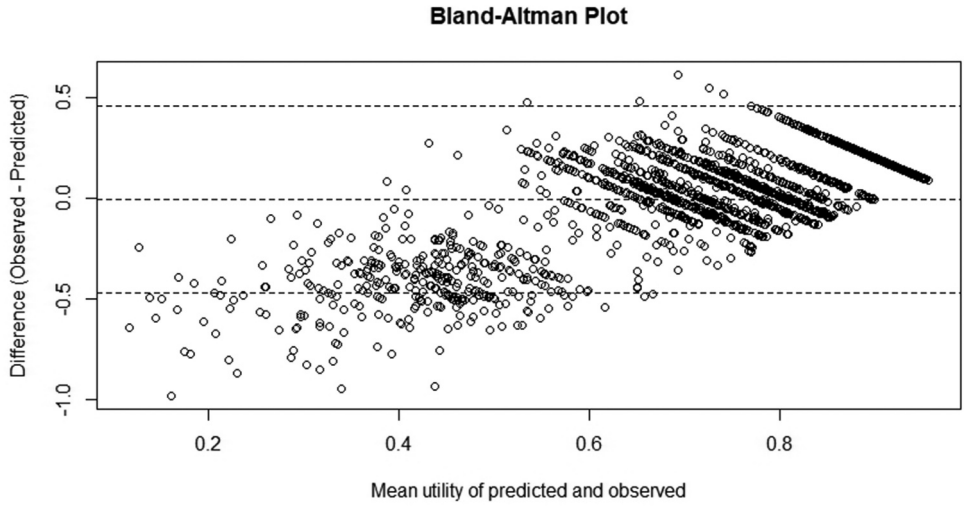

Figure 3. Bland-Altman plots of agreement between predicted and observed EQ-5D utilities using a LASSO model trained and tested on a dataset (n = 2,029) consisting of all individual HoNOS items and demographics. Observed limits range from −0.47 to 0.46, where 93.4% of observations lie within the confidence intervals.

instruments. Secondarily, the same approach was applied to develop a similar mapping algorithm using the PANSS. Data for this study was derived from a large sample of outpatient records from Dutch patients with a history of psychotic disorder. From the mapping algorithms applied in this study, the LASSO models were found to have the best predictive performance, in line with a recent report recognizing the LASSO as superior for mapping utilities compared to a broad selection of alternative statistical and machine learning methods [26]. LASSO provides regression coefficients that can be implemented to map utility scores for health economic purposes. As indicated by the R-squared and MAE of 0.261 and 0.187 for Model 3, respectively, the predictive accuracy found by our HoNOS mapping algorithm is on the lower end, which was additionally reflected in the Bland-Altman plot [46]. Based on this low to moderate predictive accuracy, use of the algorithm should be avoided when doing individual patient predictions. However, since in economic evaluations only the relative difference between treatment group averages (i.e. incremental effect) is considered, and the same regression coefficients are used to estimate utilities for both treatment groups, the predicted HRQoL values could be relevant on a group level in the absence of other HRQoL estimates. Therefore, when aiming to estimate increments between population-based EQ-5D utility scores (i.e. the difference between two group averages) to be used in economic evaluations, the algorithm could be used cautiously. Looking at the range of predicted utility scores by our best-performing model, it appears unfit for patient populations for which the mean EQ-5D utility is expected to score outside of the midrange utility scores.

Overall, predictive accuracy of our mapping algorithm can be considered moderate at best, however, similar predictive performance levels are frequently seen in other mapping algorithms for disease-specific instruments to the EQ-5D [21,24]. For example, two earlier studies using the PANSS subscales and demographics to map the EQ-5D (using the 3 L version in a mixed European study (n = 1208) [30], and the 5 L version in a Singaporean sample (n = 239) [31]), reported both slightly lower and higher R-squared values of 0.104 and 0.352, respectively. On top of that, an earlier review indicated that mapping performance is a lot more variable and frequently lower for disease-specific to generic mapping models, as compared to models mapping two generic instruments [47]. Despite this, mapping algorithms are required to inform economic evaluations under the existing lack of preference-based measures and subsequent evidence gaps, and are currently a widely accepted method by policy bodies. This is also supported by a review of appraisals by the United Kingdom's National Institute for Health and Excellence which found that mapping algorithms were used in over a quarter of appraisals to obtain EQ-5D utilities, of which the majority originated from disease-specific measures [48]. Although more conceptual overlap is always preferred, a model's predictive ability does not dictate its appropriateness for use, but rather does its applicability to the specific context of its use [49].

Despite reasonable visible conceptual overlap when comparing the HoNOS or PANSS and the EQ-5D, each of these instruments aim to serve very different purposes and perspectives, whilst using different instrumental structures. For example, the HoNOS being a clinician-reported instrument evaluating a patient's health and social functioning, whereas the EQ-5D is a self-reported instrument evaluating one's HRQoL. As a result, very strong correlation between instruments was not to be expected, and although discrepancies between the different types of instruments provide no evidence for the use or superiority of either it is, however, important to recognize how their differences and limitations may relate to the lack of observed correlation and model performance.

A first explanation may relate to a frequently recognized limitation of the EQ-5D in patients with SMIs, being its lack of responsiveness and presence of ceiling effects [11,50]. In our results ceiling effects are also clearly observed, given that as much as 25% of patients report full health despite their illness, whilst concurrently associated with a broad range of HoNOS and PANSS scores, as in line with earlier

reports using these and other mental health-specific measures [10,29]. Here, however, it is imperative to recognize the importance of subjectivity when measuring one's quality of life (using the EQ-5D) and how this is highly impacted by the personal standards and expectations set by the patient themselves, especially for concepts relating to daily living activities and occupation. Since this study sample consists of relatively older patients with an average of 22 years since disease onset, experiencing relatively little symptoms, it is likely that patients have become more adjusted to living with their disorder over time, further reducing its impact on their HRQoL measurable by the EQ-5D. Although the newer 5-level version of the EQ-5D might allow for some improvement in responsiveness, our results add to the existing discussing surrounding the inappropriateness of the EQ-5D in patients with SMIs and need for better instruments [12,50,51].

Next, it is also important to consider that various factors may introduce differences in reporting between patients and clinicians, ultimately contributing to the lack of observed correlation between instruments. First, patients and clinicians may differ in their respective perception of the patients health and functioning. Here, patients may have different insights or standards relating to their level of wellbeing and functioning, resulting in higher or lower scores, potentially also driven by the impact of their disease on self-insight and metacognition. Similarly, clinicians may also have insufficient insight in a patients current state, resulting in diverging estimates of their daily wellbeing and (social and symptomatic) functioning [52,53]. Second, patients and clinicians may differ in their interpretation of the items (i.e. understanding of the concept being evaluated) affecting their scores, or especially in case of the EQ-5D, patients may associate a wider range of disease severity with the same instrument level. In case of the HoNOS, a relatively elaborate guidance associating symptoms/impairment with respective levels is present to guide clinicians in choosing instrument levels, aiming to reduce this subjectivity.

Finally, it should be considered that the two instruments simply do not measure the same constructs, therefore lacking sufficient convergent validity. When comparing the HoNOS and EQ-5D, conceptual overlap was expected for at least 4 items of the HoNOS (5: physical illness relates to mobility and pain/discomfort, 7: depressed mood relates to anxiety/depression, 10: daily living activities relates to self-care, 12: occupation and activities relates to usual activities), and although observed correlations in our matrix were relatively higher (except for HoNOS item 12: occupation and activities), they remained moderately low. Correlations between the remaining 8 hoNOS items were largely absent (especially for HoNOS items 3, 4, 6, 11 and 12), reducing overall convergence between the instruments and, consequently, model performance. When looking at the correlations between the EQ-5D and PANSS, similar conclusions can be drawn.

This study has several strengths that should be highlighted. First of all, this study made use of a large dataset including 1,853 complete patient profiles (of which 1,164 were used to map the PANSS), supporting the validity of our results. In order to find the best-performing mapping algorithm, a machine learning approach was used combined with cross-validation to prevent overfitting. In addition to that, various model specifications were explored including demographic variables and items or subscales, as well as two additional sensitivity analyses, including diagnosis in the model specification and an alternative imputation method.

Furthermore, the results of this study should be seen in the light of some limitations. The dataset used in this study was derived from the PHAMOUS study and included the primary diagnoses at time of the assessment, as well as the PANSS, EQ-5D-3 L, and HoNOS scores, however, no information was included regarding (physical) co-morbidities of patients. Although reported in HoNOS item 5, a deeper insight in the presence of physical co-morbidities might have allowed for a better understanding and differentiation of how psychotic symptoms alone translate to EQ-5D domain scores. Second, our final study sample included 137 patients for which more than one observation was included in the final dataset, totaling 15.0% of the included records. Given the relatively small fraction of the total dataset, and the lack of impact on results found when excluding these records, no additional corrections were performed for nested individuals in our analyses. Third, the presence of heteroskedasticity was identified between HoNOS and utility scores using the Breusch-Pagan test, however, upon further inspection, residuals were found to be independent of the HoNOS scores, indicating that the heteroskedasticity itself is not skewed. Since heteroskedasticity does not affect the predictive accuracy of the model itself, but rather impacts the standard deviations of model coefficients in terms of consistency and efficiency, we decided not to perform additional corrections as this would likely not improve predictive performance [54]. However, as a result, heteroskedasticity does impact the reliability of the estimated coefficients when used for statistical inference, meaning that interpretation for clinical significance of our identified coefficients and their values should be refrained from. Finally, it is important to recognize that regression-based approaches have been shown to be susceptible for regression to the mean, in which case extreme observed values tend to be under- and overestimated. Thompson *et al.* (2017) suggest to perform linear and equipercentile equating to overcome this issue. However, given the computational complexity of this method in combination with the large number of variables included in our models, this was not deemed feasible [55]. We did, however, make use of both regressions-based (i.e. LASSO) and tree-based (i.e. RF) models to examine the impact of using different analytical approaches on our mapping problem.

5. Conclusion

To the best of our knowledge, this study presents the first mapping algorithm to translate (frequently used) clinician-rated HoNOS scores to self-reported EQ-5D-3 L utilities, as well as an alternative to map utilities using the clinician-rated PANSS, developed in the Dutch context. In this article, we have presented a thoroughly developed mapping algorithms using individual HoNOS and PANSS item scores and patient characteristics. Despite their suboptimal predictive performance, when applied to a patient population with similar characteristics, this mapping algorithm can be used

to predict population-based EQ-5D-3 L utilities for health economic purposes if no other alternatives are available. The results of this study support the frequently emphasized importance of developing disease-specific instruments which can be used in (mental) health economic evaluations.

Funding

This work was supported by a grant from the Netherlands Organization for Health Research and Development [Grant No. # 60-26 63800-98-5600]. Furthermore, this work used data collected in the PHAMOUS study, supported by the Rob Giel Research center which is structurally funded by the following mental health organizations: Lentis, GGZ Friesland, GGZ Drenthe, Mediant, Dimence Groep, and the University Center for Psychiatry of the University Medical Center Groningen.

Declaration of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Authors contributions

Overall project design and funding acquisition: J Lokkerbol, B Wijnen, S Evers. Writing of manuscript: A Kleijburg, B Wijnen. Critical appraisal of manuscript: A Kleijburg, B Wijnen, W den Hollander, S Evers, PHAMOUS investigators, H Kroon, J Lokkerbol. All authors have read and approved the manuscript.

Acknowledgments

We gratefully acknowledge all those involved with the collection of data and coordination of the PHAMOUS study, the PHAMOUS investigators for all insights provided during this study, with special recognition of Dr. E. Visser for her support with the data.

Data availability statement

Under the General Data Protection Regulation (GDPR), data of the PHAMOUS study is considered pseudonymized rather than anonymized and is therefore still regarded as personal data. Given that participants have not given informed consent to have their personal data publicly shared, PHAMOUS is legally and ethically not allowed to publish their dataset. Data is therefore only available upon request at the Rob Giel Research center (Data Science Center), e-mail e.visser03@umcg.nl.

Reviewer disclosure

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

ORCID

Anne Kleijburg http://orcid.org/0000-0001-6223-3027

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Schinnar AP, Rothbard AB, Kanter R, et al. An empirical literature review of definitions of severe and persistent mental illness. Am J Psychiatry. 1990;147(12):1602–1608.
2. Delespaul P. Consensus regarding the definition of persons with severe mental illness and the number of such persons in the Netherlands [Dutch article]. Tijdschr Psychiatr. 2013;55(6):427–438.
3. OECD/EU. Health at a glance: Europe 2018: state of health in the EU cycle [internet]. Paris: OECD Publishing; 2018 [cited 2022 Sep 30]. Available from: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2018_health_glance_eur-2018-en
4. National Institute for Public Health and the Environment. Cost of illness 2019 [internet]. 2019 [cited 2022 Sep 30]. Available from: https://statline.rivm.nl/#/RIVM/nl/dataset/50091NED/table?ts=1660223783679
5. König H, König HH, Gallinat J, et al. Excess costs of mental disorders by level of severity. Soc Psychiatry Psychiatr Epidemiol [Internet]. 2022 [cited 2022 Nov 16];1:1–13.
6. National Institute for Health and Care Excellence (NICE). The reference case: Guide to the methods of technology appraisal. 2013 [cited 2022 Dec 20].Available from: https://www.nice.org.uk/process/pmg9/chapter/the-reference-case
7. Dutch Institute National Health Care (Zorginstituut Nederland). Dutch guidelines for economic evaluations in healthcare. Diemen: Zorginstituut Nederland; 2015.
8. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. Br Med Bull. [Internet]. 2010 [cited 2022 Nov 1];96(1):5–21. doi: 10.1093/bmb/ldq033
9. Feng YS, Kohlmann T, Janssen MF, et al. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. Qual Life Res. [Internet]. 2021 [cited 2022 Nov 1];30(3):647–673. doi: 10.1007/s11136-020-02688-y
10. Brazier J, Connell J, Papaioannou D, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technol Assess [Internet]. 2014 [cited 2022 Oct 29];18(34):1–188. doi: 10.3310/hta18340
11. Payakachat N, Ali MM, Tilford JM. Can the EQ-5D detect meaningful change? A systematic review. Pharmacoeconomics [Internet]. 2015 [cited 2022 Sep 30];33(11):1137. doi: 10.1007/s40273-015-0295-6
12. Mulhern B, Mukuria C, Barkham M, et al. Using generic preference-based measures in mental health: psychometric validity of the EQ-5D and SF-6D. Br J Psychiatry [Internet]. 2014 [cited 2022 Oct 29];205(3):236–243. doi: 10.1192/bjp.bp.112.122283
13. Karow A, Wittmann L, Schöttle D, et al. The assessment of quality of life in clinical practice in patients with schizophrenia. Dialogues Clin Neurosci [Internet]. 2014 [cited 2022 Oct 29];16(2):185. doi: 10.31887/DCNS.2014.16.2/akarow
14. Wilkinson G, Hesdon B, Wild D, et al. Self-report quality of life measure for people with schizophrenia: the SQLS. Br J Psychiatry [Internet]. 2000 [cited 2023 Oct 31];177(1):42–46. doi: 10.1192/bjp.177.1.42
15. Kay SR, Flszbeln A, Qpjer LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull. 1967 [cited 2023 Nov 29];13(2):261–276. doi: 10.1093/schbul/13.2.261
16. Citrome L, Mychaskiw MA, Cortez A, et al. Clinical outcome assessment instruments in schizophrenia: a scoping literature review with a focus on the potential of patient-reported outcomes. Innov Clin Neurosci [Internet]. 2023 [cited 2023 Oct 31];20(4–6):14.
17. National Institute for Health and Care Excellence (NICE). NICE health technology evaluations: the manual [Internet]. 2022 [cited 2022 Jan 1]. Available from: https://www.nice.org.uk/process/pmg36/chapter/introduction-to-health-technology-evaluation
18. Von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton: Princeton University Press; 1944.
19. Drummond MF, Sculpher MJ, Torrance GW, et al. Methods for the economic evaluation of health care programme. Third ed. [Internet]. Oxford University Press; 2005 [cited 2023 Apr 22: Available from: https://pure.york.ac.uk/portal/en/publications/methods-for-the-economic-evaluation-of-health-care-programme-thir
20. Brazier J, Ratcliffe J, Saloman J, et al. Measuring and Valuing Health Benefits for Economic Evaluation. 2nd ed. Oxford: Oxford Academic; 2016. [cited 2022 Jan 29].

21. Wijnen BFM, Mosweu I, Majoie MHJM, et al. A comparison of the responsiveness of EQ-5D-5L and the QOLIE-31P and mapping of QOLIE-31P to EQ-5D-5L in epilepsy. Eur J Health Econ [Internet]. 2018 [cited 2021 Nov 15];19(6):861–870. doi: 10.1007/s10198-017-0928-0
22. Lim J, Choi SE, Bae E, et al. Mapping analysis to estimate EQ-5D utility values using the COPD assessment test in Korea. Health Qual Life Outcomes [Internet]. 2019 [cited 2021 Nov 15];17(1):1–7. doi: 10.1186/s12955-019-1148-3
23. Teckle P, McTaggart-Cowan H, Van der Hoek K, et al. Mapping the FACT-G cancer-specific quality of life instrument to the EQ-5D and SF-6D. Health Qual Life Outcomes [Internet]. 2013 [cited 2022 Sep 23];11(1):1–10. doi: 10.1186/1477-7525-11-203
24. Kularatna S, Senanayake S, Chen G, et al. Mapping the Minnesota living with heart failure questionnaire (MLHFQ) to EQ-5D-5L in patients with heart failure. [cited 2021 Nov 15]. doi: 10.1186/s12955-020-01368-2
25. Longworth L, Rowen D. Mapping to obtain EQ-5D utility values for use in nice health technology assessments. Value Heal [Internet]. 2013;16(1):202–210. doi: 10.1016/j.jval.2012.10.010

•• **Comprehensive review and guidance on mapping the EQ-5D.**

26. Aghdaee M, Parkinson B, Sinha K, et al. An examination of machine learning to map non-preference based patient reported outcome measures to health state utility values. Heal Econ (UK). 2022;31(8):1525–1557. doi: 10.1002/hec.4503

•• **Comprehensive comparison of machine learning techniques for mapping the EQ-5D.**

27. Speak B, Muncer S. The structure and reliability of the health of the nation outcome scales. Australas Psychiatry. 2015;23(1):66–68. doi: 10.1177/1039856214563851
28. Lovaglio PG, Monzani E. Health of the nation outcome scales evaluation in a community setting population. doi: 10.1007/s11136-011-0071-9
29. Franklin M, Mukuria C, Mulhern B, et al. Measuring the burden of schizophrenia using clinician and patient-reported measures: an exploratory analysis of construct validity. Patient - patient-centered outcomes res [Internet]. The Patient - Patient Centered Outcomes Res. 2019 [cited 2023 Oct 3];12(4):405–417. doi: 10.1007/s40271-019-00358-x

• **Exploratory analysis on a variety of schizophrenia and utlity measures.**

30. Siani C, de Peretti C, Millier A, et al. Predictive models to estimate utility from clinical questionnaires in schizophrenia: findings from EuroSC. Qual Life Res. [Internet]. 2016 [cited 2024 Jan 16];25(4):925. doi: 10.1007/s11136-015-1120-6
31. Abdin E, Chong SA, Seow E, et al. Mapping the positive and negative syndrome scale scores to EQ-5D-5L and SF-6D utility scores in patients with schizophrenia. Qual Life Res [Internet]. 2019 [cited 2024 Jan 3];28(1):177–186. doi: 10.1007/s11136-018-2037-7
32. Bartels-Velthuis AA, Visser E, Arends J, et al. Towards a comprehensive routine outcome monitoring program for people with psychotic disorders: the pharmacotherapy monitoring and outcome survey (PHAMOUS). Schizophr Res [Internet]. 2018;197:281–287. doi: 10.1016/j.schres.2018.01.016
33. Wing JK, Beevor AS, Curtis RH, et al. Health of the nation outcome scales (HoNOS): research and development. Br J Psychiatry. 1998;172(1):11–18. doi: 10.1192/bjp.172.1.11
34. Parabiaghi A, Kortrijk HE, Mulder CL. Defining multiple criteria for meaningful outcome in routine outcome measurement using the health of the nation outcome scales. Soc Psychiatry Psychiatr Epidemiol [Internet]. 2014 [cited 2023 Oct 3];49(2):291–305. doi: 10.1007/s00127-013-0750-7
35. Kay SR, Opler LA, Lindenmayer JP. Reliability and validity of the positive and negative syndrome scale for schizophrenics. Psychiatry Res. 1988;23(1):99–110. doi: 10.1016/0165-1781(88)90038-8
36. Brooks R, De Charro F. EuroQol: the current state of play. Health Policy (NY). 1996;37(1):53–72. doi: 10.1016/0168-8510(96)00822-6
37. Rabin R, De Charro F. EQ-5D: a measure of health status from the EuroQol group. Ann Med. 2001;33(5):337–343. doi: 10.3109/07853890109002087
38. Lamers LM, McDonnell J, Stalmeier PFM, et al. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006;15(10):1121–1132. doi: 10.1002/hec.1124
39. Ratner B. The correlation coefficient: its values range between +1/−1, or do they? J Target Meas Anal Mark [Internet]. 2009 [cited 2023 Feb 10];17(2):139–142. doi: 10.1057/jt.2009.5
40. Kuhn M. The caret package. In: r found stat comput Vienna [internet]. Vienna: In: R Found Stat Comput; 2012. Available from: https//cranr-projectorg/package=caret
41. Aghdaee M, Parkinson B, Sinha K, et al. An examination of machine learning to map non-preference based patient reported outcome measures to health state utility values. Health Econ [Internet]. 2022 [cited 2022 Oct 30];31(8):1525–1557. doi: 10.1002/hec.4503
42. James G, Witten D, Hastie T, et al. An introduction to statistical learning: with applications in R. 2nd ed. New York: Springer; 2021.
43. R Core Team. R: a language and environment for statistical computing. Vienna AU: R Foundation for Statistical Computing; 2020.
44. Kyung M, Gilly J, Ghoshz M, et al. Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal. 2010;5(2):369–412. doi: 10.1214/10-BA607
45. Versteegh M, Vermeulen M, K MAA, et al. Dutch tariff for the five-level version of EQ-5D. Value Heal. 2016. [cited 2021 Apr];19(4):343–352. doi: 10.1016/j.jval.2016.01.003
46. Moore DS, Notz Maf WI. The basic practice of statistics. 6th ed. New York (NY): W.H. Freeman and Company; 2013.
47. Brazier JE, Yang Y, Tsuchiya A, et al. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ [Internet]. 2010 [cited 2021 Nov 15];11(2):215–225. doi: 10.1007/s10198-009-0168-z
48. Tosh JC, Longworth LJ, George E. Utility values in national institute for health and clinical excellence (NICE) technology appraisals. Value Heal. 2011;14(1):102–109. doi: 10.1016/j.jval.2010.10.015
49. Mukuria C, Rowen D, Harnan S, et al. An updated systematic review of studies mapping (or cross-walking) measures of health-related quality of life to generic preference-based measures to generate utility values. Appl Health Econ Health Policy. 2019 [cited 2022 Dec 20];17(3):295–313. doi: 10.1007/s40258-019-00467-6
50. Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. Value Heal. 2011;14(6):907–920. doi: 10.1016/j.jval.2011.04.006

•• **Comprehensive review on validity and responsiveness of generic utility measures in schizophrenia.**

51. Janssen MF, Bonsel GJ, Luo N. Is EQ-5D-5L better than EQ-5D-3L? A head-to-head comparison of descriptive systems and value sets from seven countries. Pharmacoeconomics [Internet]. 2018 [cited 2022 Dec 14];36(6):675. doi: 10.1007/s40273-018-0623-8
52. Nair A, Palmer EC, Aleman A, et al. Relationship between cognition, clinical and cognitive insight in psychotic disorders: a review and meta-analysis. 2013 [cited 2022 Oct 30]. doi: 10.1016/j.schres.2013.11.033
53. Takeuchi H, Fervaha G, Remington G. Reliability of a patient-reported outcome measure in schizophrenia: results from back-to-back self-ratings. Psychiatry Res. 2016;244:415–419. doi: 10.1016/j.psychres.2016.07.055
54. Wooldridge JM. Introductory econometrics: a Modern approach. Fifth. Toler. Anal. Electron. Circuits using MATHCAD. Mason (OH): South Western Cengage Learning; 2012.
55. Thompson NR, Lapin BR, Katzan IL. Mapping PROMIS global health items to EuroQol (EQ-5D) utility scores using linear and equipercentile equating. Pharmacoeconomics. 2017;35(11):1167–1176. doi: 10.1007/s40273-017-0541-1