Assessment of outcome measures for cost–utility analysis in depression: mapping depression scales onto the EQ-5D-5L

Thor Gamst-Klaussen, Admassu N. Lamu, Gang Chen and Jan Abel Olsen

Background

Many clinical studies including mental health interventions do not use a health state utility instrument, which is essential for producing quality-adjusted life years. In the absence of such utility instrument, mapping algorithms can be applied to estimate utilities from a disease-specific instrument.

Aims

We aim to develop mapping algorithms from two widely used depression scales; the Depression Anxiety Stress Scales (DASS-21) and the Kessler Psychological Distress Scale (K-10), onto the most widely used health state utility instrument, the EQ-5D-5L, using eight country-specific value sets.

Method

A total of 917 respondents with self-reported depression were recruited to describe their health on the DASS-21 and the K-10 as well as the new five-level version of the EQ-5D, referred to as the EQ-5D-5L. Six regression models were used: ordinary least squares regression, generalised linear models, beta binomial regression, fractional logistic regression model, MM-estimation and censored least absolute deviation. Root mean square error, mean absolute error and r^2 were used as model performance criteria to select the optimal mapping function for each country-specific value set.

Results

Fractional logistic regression model was generally preferred in predicting EQ-5D-5L utilities from both DASS-21 and K-10. The only exception was the Japanese value set, where the beta binomial regression performed best.

Conclusions

Mapping algorithms can adequately predict EQ-5D-5L utilities from scores on DASS-21 and K-10. This enables disease-specific data from clinical trials to be applied for estimating outcomes in terms of quality-adjusted life years for use in economic evaluations.

Declaration of interest

None.

Keywords

Statistical methodology; cost-effectiveness; EQ-5D-5L; mapping; DASS-21, K-10.

When comparing the effectiveness of competing healthcare programmes across disease areas, there is a growing interest in estimating health outcomes on a generic metric, such as quality-adjusted life years (QALYs). To enable QALY calculations, a preference-based health-related quality of life instrument, also referred to as a health state utility (HSU) instrument,[1] is essential. Such HSU instruments consist of a descriptive system and a predetermined value set that reflects the preferences of the general population, which assign a value – or utility – to each possible combination of health states in the descriptive system.

In clinical trials, however, we find condition-specific instruments to be more commonly applied than generic instruments. This is because clinicians have an affinity to the gold standard instruments within their speciality, but also because condition-specific instruments tend to identify disease-specific changes in health that might not be identified by a generic descriptive system. In cases where condition-specific data have been collected and decision makers want effectiveness to be expressed on a generic metric, there is a need for a mapping algorithm to convert condition-specific data to HSU.[1,2] Such mapping algorithms are commonly developed by distributing both measures of interest to the same respondents, and applying statistical methods to predict utilities from scores on a source instrument.

Health outcome measures

Depression is a common mental disorder and one of the main causes of disability worldwide.[3] It can last for long periods or re-occur, impairing work or school performance and the ability to cope with daily life. Although a wide range of mental health outcome measures are suitable to measure its effect, they do not produce utilities. The Depression Anxiety Stress Scales (DASS-21)[4] and Kessler Psychological Distress Scale (K-10)[5] are two of the most widely used mental health-specific instruments, assessing core symptoms of depression, anxiety and stress.[6]

The most widely used HSU instrument is the EQ-5D. A recent review supported its dominant position by revealing that 70% of cost–utility studies had applied the EQ-5D.[7] One reason for its widespread use is that it has been recommended by the National Institute for Health and Care Excellence (NICE) in the UK.[8] Studies generating mapping algorithms for producing EQ-5D utilities are increasing in number, especially after NICE endorsed mapping if the direct measure of EQ-5D utility is unavailable.[2]

This paper has three aims. First, we aim to replace the existing mapping algorithms for DASS-21 and K-10 that were recently published in the *British Journal of Psychiatry*.[6] The paper by Mihalopoulos *et al* was based on an interim EQ-5D-5L value set,[9] which was developed based on the value set for the three-level version.[10] Most recently, eight country-specific value sets have been

published for the EQ-5D-5L instrument, including four Western countries (England, the Netherlands, Spain and Canada), three Asian countries (China, Japan and Korea) and one South American (Uruguay).[11–18] The previously published mapping algorithm is already becoming obsolete in the literature after the publication of the directly elicited EQ-5D-5L official value sets.

Second, we aim to investigate if mapping algorithms for the two mental health instruments differ across countries, depending on country-specific health state preferences. Because health state preferences differ across countries,[19] their EQ-5D-5L value sets differ accordingly. Hence, there is a need to develop country-specific mapping algorithms.

Third, we aim to make important methodological contributions. Although the paper by Mihalopoulos *et al* applied two different mapping models (ordinary least squares regression (OLS) and generalised linear models (GLM)),[4] this paper further investigates the relative merit of six regression models. Best practice for reporting mapping studies are followed, based on the Mapping Preference-based Measures Reporting Standards statement.[20]

Method

Sample

Data were obtained from the Multi-Instrument Comparison study, which is based on an online survey administered in six countries (Australia, Canada, Germany, Norway, UK and USA) by a global panel company, CINT Australia Pty Ltd.[21] The current paper is based on respondents who were diagnosed with depression ($n =$ 917). The depression group were asked to describe their condition on both the DASS-21 and the K-10, as well as the EQ-5D-5L. For further details on respondent description, see Richardson *et al*[21] and Mihalopoulos *et al.*[6]

Instruments

DASS-21

The DASS-21 comprises 21 items, each with a four-point severity scale indicating how much the statement applies to the respondent (did not apply to me; applied to some degree; applied a considerable degree; applied very much or most of the time).[4] It comprises three seven-item subscales that measure core symptoms of depression, anxiety and stress. The items of each subscale are summed into a scale score ranging from 0 to 42, where lower values indicate fewer problems.

K-10

The K-10 measures psychological distress comprising 10 items asking about anxiety and depressive symptoms experienced in the past 4 weeks.[5] Each item has five response levels (none of the time; a little of the time; some of the time; most of the time; all of the time). Items are summed into a scale score of 10–50, where lower values indicate less problems.

EQ-5D-5L

The EQ-5D consists of five items/dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The five-level version (EQ-5D-5L) is based on the original three-level version (EQ-5D-3L) by inserting two more response levels to each dimension to reduce potential ceiling effects and improve reliability and sensitivity.[22] The five response levels are no problem, slight problem, moderate problem, severe problem and unable to/extreme problem. The instrument produces 3125 (5^5) health states. The utility scores were calculated by applying eight

country-specific value sets: England, the Netherlands, Spain, Canada, China, Japan, Korea and Uruguay.[11–18]

Statistical analysis

Descriptive

Spearman's rank correlation (r_s) and exploratory factor analyses (EFA) were used to assess the degree of conceptual overlap between the source instruments (DASS-21 and K-10) and the target instrument (EQ-5D-5L). EFA with principal axis factoring was used, which has been recommended as the preferred method of factor extraction.[23] An eigenvalue >1 and the scree test was used as selection criteria to extract underlying constructs.[23] Further, as the extracted factors are usually correlated,[24] a promax rotation was applied.[25] Correlations between the extracted factors were also observed (see supplementary Table 2a and b).

A direct mapping technique was applied by regressing EQ-5D-5L utility index onto the source instrument, either the DASS-21 subscale scores or K-10 total score. Six alternative models were estimated and compared (as described below). For every regression model, a forward stepwise selection method was used for variable selection ($P < 0.05$). To make mapping equations applicable to all data-sets, only age and gender were considered as covariates. Interaction and squared terms were only considered if the original variable was significant. Indirect mapping (i.e. response mapping) is not suitable in this case because of the limited overlap between the two depression scales and the EQ-5D-5L. This issue is demonstrated in the EFA results. In indirect mapping, responses to each of the five dimensions of the EQ-5D-5L will be predicted in the first step before further applying the country-specific value sets. With limited overlap across dimensions in two instruments (i.e. mainly mental health dimension in EQ-5D-5L), the prediction error for four physical health-related dimensions of EQ-5D-5L will be large.

Regression models

OLS is the most commonly used regression model in mapping studies,[26] and requires data to be normally distributed with constant variance. Unlike the OLS, the GLM allows for skewed distribution (i.e. non-normal distribution) of the dependent variable. Gamma family and log-link function fit the model well for GLM in this data. Because gamma and log function are defined for non-negative values, EQ-5D-5L disutility (where disutility is equal to 1 – EQ-5D-5L utility) was used. Beta binomial regression allows the dependent variable to be skewed and is capable of modelling bounded dependent variables restricted between 0 and 1, which is often the case with utility instruments. As this parametric model is not defined at the boundary values, the outcome values should be restricted to a 0–1 range, excluding 0 and 1. This can be achieved by linear transformation $[Y(N-1) + 0.5]/N$ following earlier literature,[27] where N refers to sample size, and Y is the dependent variable. For applications of the beta binominal regression model, see Khan *et al*[28] for detail. Another similar approach for modelling bounded data defined on [0, 1] scale that involves a semi-parametric approach is the fractional regression model (FRM). It was developed to address the modelling of empirical bounded dependent variables, such as proportions and percentages, that exhibit piling-up at one of the two corners.[29] In the FRM model, EQ-5D-5L scores are linearly transformed onto a 0–1 scale by subtracting the minimum score from EQ-5D-5L and then dividing by the range. For both beta binomial and FRM, the logit link function fits the model well in this data and is applied here. The logit transformation used in the prediction of EQ-5D-5L utility is given as:

$$\frac{\exp(\beta X)}{1 + \exp(\beta X)},$$

where X is a vector of predictors (i.e., the DASS depression and anxiety subscales score or the K-10 overall score) and age, and β is a vector of estimated coefficients.

MM-estimation is a robust regression estimation approach that is appropriate when the residual distribution is non-normal or some outliers affect the model.[30] MM-estimation estimates the regression parameter by S-estimation, which minimise the scale of the residual from M-estimation and then proceeds with M-estimation. The S in S-estimation stands for the scale of the residual, the M in M-estimation stands for maximum likelihood type and the MM in MM-estimation stands for minimising M-estimation.[30] It aims to obtain estimates that have a high breakdown value and is more efficient. The breakdown value is a common measure of the proportion of outliers that can be addressed before these observations affect the model.[31] Censored least absolute deviations (CLAD) model is more appropriate for outcome variables censored at one or both end-points.[32] The CLAD model is a semi-parametric estimator that is robust to distributional assumptions and heteroscedasticity because it uses median values rather than means among similar groups, as medians are likely to be less affected by censoring.

Model performance

In line with previous research,[26] the predictive performance of each model described above was assessed by mean absolute error (MAE) and root mean square error (RMSE). Both were computed for the full sample (where lower values indicate better fit). The MAE is defined as the average of absolute difference between observed and predicted EQ-5D-5L. The RMSE is the square root of the average of the squared differences between observed and predicted EQ-5D-5L. Both MAE and RMSE were adjusted for the degrees of freedom, as the number of independent variables may differ across models.

It has been shown that the wider the scale length of the EQ-5D-5L, the larger the error.[33] Therefore, adjusting for scale differences would allow reasonable comparison between data-sets or models with different scales. Although there are no standard ways of normalisation in the literature, we normalise both MAE and RMSE to the range (defined as the difference between the maximum and the minimum values) of the measured data. Such normalised RMSE (NRMSE) and normalised MAE (NMAE) are non-dimensional and enable us to compare data-sets and models with different units or scales. Lastly, the performance of each model was also assessed by the square of the correlation coefficient between the observed and predicted values adjusted for the number of predictors in the model (adjusted r^2).[34] In addition, binned scatter plots between the observed and predicted EQ-5D-5L utilities were reported to visualise the predictive performance of each model.

To investigate the generalisability of the preferred mapping algorithms, cross-validation was performed by splitting the existing data into two: estimation and validation samples via random selection procedures. In this study, the total sample was randomly divided into two equal groups to evaluate the model fit in out-of-sample data. The model was fitted on the estimation sample, and the resulting parameters from the fitted model were then used to predict the EQ-5D-5L on the validation sample. This procedure has been repeated by reversing the validation and estimation sample. The average RMSE, MAE and r^2 for both iterations were calculated for comparison of the models' predictive performance. Lastly, the best-fitting model was estimated with the full sample ($N = 917$). All statistical analyses were conducted with Stata version 14.2 (StataCorp, College Station, Texas, USA), except the EFA, which was carried out in SPSS version 24 (IBM Corp, Armonk, New York, USA).

Ethical approval

Data for this study were obtained from the Multi-Instrument Comparison project, which was approved by the Monash University Human Research Ethics Committee (numbers CF11/1758-2011000974 and CF11/3192-2011001748).

Results

Sample characteristics are presented in Table 1. The estimated EQ-5D-5L utility scores varied both in the mean score and the range, depending on the choice of country-specific value sets. In the depression sample, the mean EQ-5D-5L utility ranged from 0.59 (Dutch value set) to 0.83 (Uruguayan value set). The minimum utility score ranged from −0.41 in the Dutch value set to 0.12 in the Korean and Uruguayan value set. Spearman's rank correlations are presented in supplementary Table 1, available at https://doi.org/10.1192/bjo.2018.21. Among EQ-5D-5L dimensions, anxiety/depression dimension produced the highest correlation with the source instruments ($r_s \geq 0.50$), whereas mobility dimension produced the lowest ($r_s \leq 0.25$). The three DASS-21 subscales were highly correlated with each other ($r_s = 0.63–0.73$).

The EFA was appropriate as indicated by a Kaiser–Meyer–Olkin measure of sampling adequacy of >0.90 and a highly significant Bartlett's test of sphericity. The pattern matrix for EFA with at least 0.30 (factor) loadings are reported in Table 2a and b. The EFA analysis for DASS-21 and EQ-5D-5L items produced four underlying factors (depression, anxiety, stress and physical functioning), explaining 60% of the variance. The extracted factors replicate the original factor structure of DASS-21 subscales: depression, anxiety and stress, except item 2: 'I was aware of dryness of my mouth', which was originally part of the anxiety subscale. However, this item produced weak loadings on three factors: physical (0.288), stress (0.197) and anxiety (0.161). The result revealed conceptual overlap between the anxiety/depression dimension of EQ-5D-5L and the extracted DASS-21 depression factor. All remaining (four) EQ-5D-5L dimensions were mainly loaded on the fourth factor (i.e. physical functioning).

Considering the result with K-10 items, three factors were extracted: depression, anxiety and physical functioning (Table 2b). Again, only EQ-5D-5L anxiety/depression dimension loaded on the extracted K-10 depression factor. No single item from K-10 items was mainly loaded to the last factor (physical), which was formed by the first four dimensions of EQ-5D-5L. The structure matrix presented in supplementary Table 2a and b (which shows the correlation of each item with the extracted factors) revealed

Table 1	Sample characteristics ($N = 917$)		
Characteristic	Mean (s.d.)	Minimum	Maximum
Gender, N (%)			
Male	313 (34.1)		
Female	604 (65.9)		
Age, years	42.02 (13.38)	18	90
DASS-21			
Depression	21.02 (11.61)	0	42
Anxiety	13.20 (9.93)	0	42
Stress	19.54 (10.07)	0	42
K-10	29.19 (8.55)	10	50
EQ-5D-5L utilities			
Canada	0.69 (0.21)	0.0001	0.95
England	0.69 (0.22)	−0.17	1
the Netherlands	0.59 (0.27)	−0.41	1
Spain	0.66 (0.20)	−0.14	1
China	0.67 (0.24)	−0.25	1
Japan	0.68 (0.16)	0.10	1
Korea	0.71 (0.16)	0.12	1
Uruguay	0.83 (0.15)	0.12	1

Table 2a Exploratory factor analysis – pattern matrix				
DASS-21 items	Factor			
	Depression	Anxiety	Stress	Physical
1. I found it hard to wind down.			0.519	
2. I was aware of dryness of my mouth.				[0.288]
3. I couldn't seem to experience any positive feeling at all.	0.707			
4. I experienced breathing difficulty (e.g. excessively rapid breathing, breathlessness in the absence of physical exertion).		0.505		
5. I found it difficult to work up the initiative to do things.	0.481			
6. I tended to overreact to situations.			0.725	
7. I experienced trembling (e.g. in the hands).		0.633		
8. I felt that I was using a lot of nervous energy.		0.648		
9. I was worried about situations in which I might panic and make a fool of myself.		0.626		
10. I felt that I had nothing to look forward to.	0.879			
11. I found myself getting agitated.			0.677	
12. I found it difficult to relax.			0.531	
13. I felt downhearted and blue.	0.743			
14. I was intolerant of anything that kept me from getting on with what I was doing.			0.552	
15. I felt I was close to panic.		0.680		
16. I was unable to become enthusiastic about anything.	0.799			
17. I felt I was not worth much as a person.	0.823			
18. I felt that I was rather touchy.			0.584	
19. I was aware of the action of my heart in the absence of physical exertion (e.g. sense of heart rate increase, heart missing a beat).		0.630		
20. I felt scared without any good reason.		0.766		
21. I felt that life was meaningless.	0.886			
EQ-5D-5L items				
1. Mobility				0.872
2. Self-care				0.571
3. Usual activities				0.725
4. Pain/discomfort				0.703
5. Anxiety/depression	0.445			

Note. Loadings below 0.30 not shown, except for item two of the Depression Anxiety Stress Scales (DASS-21), where the highest loading is reported in brackets. Rotation method: promax with Kaiser normalisation.

similar results with Spearman's correlation coefficients (supplementary Table 1).

Table 3 presents model performance based on the English value set. Fractional logistic regression performed best when we consider adjusted r^2 and NRMSE for both DASS-21 and K-10. In terms of NMAE, CLAD and MM-estimation performed best for DASS-21, and MM-estimation performed best for K-10. Similar result was revealed by cross-validation. This result was also supported by the scatter plot (supplementary Fig. 1).

Model performance based on other country specific value sets are presented in supplementary Table 3a and b. Except for the Japanese value set, FRM was preferred in terms of adjusted r^2 and NRMSE, whereas MM-estimation or CLAD was preferred with NMAE. For the Japanese value set, beta binomial regression was a preferred model when NMAE and NRMSE were considered, whereas FRM was preferred in terms of adjusted-r^2.

Table 4 presented regression results when the English value set was applied. Based on the criteria described above, best-fitting regression results for the other country-specific value sets were presented in supplementary Table 4. When DASS-21 was the source instrument, the depression and anxiety subscales and age were significant ($P < 0.05$) predictors in all models. When K-10 was the

Table 2b Exploratory factor analysis – pattern matrix			
K-10 items	Factor		
	Depression	Anxiety	Physical
1. In the past 4 weeks, about how often did you feel tired for no good reason?	0.435		
2. In the past 4 weeks, about how often did you feel nervous?		0.551	
3. In the past 4 weeks, about how often did you feel so nervous that nothing could calm you down?		0.618	
4. In the past 4 weeks, about how often did you feel hopeless?	0.856		
5. In the past 4 weeks, about how often did you feel restless or fidgety?		0.794	
6. In the past 4 weeks, about how often did you feel so restless that you could not sit still?		0.878	
7. In the past 4 weeks, about how often did you feel depressed?	0.954		
8. In the past 4 weeks, about how often did you feel that everything was an effort?	0.689		
9. In the past 4 weeks, about how often did you feel so sad that nothing could cheer you up?	0.772		
10. In the past 4 weeks, about how often did you feel worthless?	0.872		
EQ-5D-5L items			
1. Mobility			0.871
2. Self-care			0.583
3. Usual activities			0.713
4. Pain/discomfort			0.675
5. Anxiety/depression	0.650		

Note. Loadings below 0.30 are not shown. Rotation method: promax with Kaiser normalisation.
K-10, Kessler Psychological Distress Scale.

Table 3 Comparison of model performance based on English value set for the EQ-5D-5L

Model	DASS-21						K-10					
	Full sample estimation			Cross-validation			Full sample estimation			Cross-validation		
	adj. r^2	NMAE	NRMSE	adj. r^2	NMAE	NRMSE	adj. r^2	NMAE	NRMSE	adj. r^2	NMAE	NRMSE
OLS	0.3320	0.1145	0.1539	**0.3343**	0.1146	0.1552	0.3288	0.1135	0.1543	0.3276	0.1147	0.1545
GLM	0.3324	0.1139	0.1541	0.3293	0.1135	0.1541	0.3285	0.1130	0.1544	0.3208	0.1126	0.1543
Beta binomial	0.3380	0.1159	0.1536	0.3334	0.1157	0.1552	**0.3345**	0.1150	0.1541	**0.3336**	0.1155	0.1560
FRM	**0.3387**	0.1135	**0.1532**	0.3309	0.1136	**0.1531**	**0.3345**	0.1125	**0.1536**	0.3324	0.1124	**0.1541**
MM-estimation	0.3318	**0.1121**	0.1575	0.3295	**0.1121**	0.1565	0.3287	**0.1111**	0.1574	0.3244	**0.1108**	0.1574
CLAD	0.3306	**0.1121**	0.1567	0.3262	0.1124	0.1566	0.3288	**0.1111**	0.1577	0.3283	0.1118	0.1607

Note. The best results are in bold type.
adj. r^2, square of correlation coefficient between predicted and observed EQ-5D-5L, penalised for number of predictors; CLAD, censored least absolute deviation; DASS-21, Depression Anxiety Stress Scales; FRM, fractional regression model; GLM, generalised linear model; K-10, Kessler Psychological Distress Scale; NMAE, normalised mean absolute error; NRMSE, normalised root mean square error; OLS, ordinary least squares regression.

Table 4 Best-fitting regression results predicting EQ-5D-5L utilities[a] from DASS-21 and K-10

	Coefficient[b]	95% CI	
	(Standard error)	Lower	Upper
	DASS-21		
DASS-Depression	−0.0236 (0.0027)	−0.0290	−0.0183
DASS-Anxiety	−0.0320 (0.0035)	−0.0389	−0.0251
Age	−0.0132 (0.0020)	−0.0171	−0.0091
Constant	2.5190 (0.1040)	2.3152	2.7228
	K-10		
K-10	−0.06476 (0.00337)	−0.0714	−0.0582
Age	−0.01382 (0.00202)	−0.0178	−0.0099
Constant	3.52220 (0.13543)	3.2562	3.7882

Note. Robust standard errors are shown in parentheses.
DASS-21, Depression Anxiety Stress Scales; K-10, Kessler Psychological Distress Scale.
a. Based on the English value set.
b. All coefficients significant at $P < 0.001$.

source instrument, the K-10 total scale and age were significant ($P < 0.05$) predictors.

Unlike the linear regression model, the beta binomial and FRM estimation produce non-linear relationships between predictors and the targeting EQ-5D-5L utilities. The beta binomial and FRM coefficients are not directly interpretable. In this study, we are not interested in interpretation of the raw coefficients but rather in the prediction of EQ-5D-5L utilities. An example has been given below to show how to use the results reported in Table 4 to calculate the predicted EQ-5D-5L utilities from K-10, using the logit transformation. Assuming the mean value for both age and the K-10 score (i.e. 42 and 29.2, respectively), the predicted EQ-5D-5L utility can be calculated as $Y = \exp(3.52220 - 0.01382 \times 42 - 0.06476 \times 29.2) / (1 + \exp(3.52220 - 0.01382 \times 42 - 0.06476 \times 29.2)) = 0.741$.

Discussion

Given the increasing use of the EQ-5D instrument in healthcare decision-making, there is a need for updated mapping of disease-specific instruments onto the recently developed preference-based value sets for the new 5L version of the EQ-5D. This study aimed at developing mapping algorithms from two widely used depression rating scales, the DASS-21 and the K-10, onto eight official country-specific EQ-5D-5L value sets. Further, we assessed the merits of six different regression models.

Based on the comparison of these regression models, the result showed that the FRM model was generally the best performing model in predicting the EQ-5D-5L utility index. The only exception was for the Japanese value set, where the beta binomial regression model was preferred. The relative performance of different

regression models was the same when either DASS-21 or K-10 was the source instrument.

In general, beta binomial regression produced the second best adjusted r^2 estimate in all cases, whereas the MM-estimation or CLAD overall produced the lowest MAE. Censoring is not a problem in our sample, where <2% report full health on EQ-5D-5L. The novelty of the FRM and the beta binomial model is that they are more appropriate for data that is bounded (as is the case for EQ-5D) and the non-linearity in the data is accounted for. Further, FRM does not make any distributional assumption about an underlying structure used to obtain the dependent variable.[29] Note that both mean and median regressions were assessed in our study. The main concern when assessing mapping results is the accuracy of the predictions. Thus, the use of mean or median regressions were the means to the end; that is, to obtain better prediction of individual utilities, which is important for cost-effectiveness analyses.

Previously, one study has published mapping equations from DASS-21 and K-10 onto EQ-5D-5L with the same data-set.[6] However, our study provides important contributions. First, the previous study only considered OLS and GLM, whereas we have compared six different regression models suitable for the sample data, e.g. problems of normality and heterogeneity of variance. Second, the previous study applied an interim value set that is already becoming obsolete after the publication of country-specific value sets that are based on directly elicited EQ-5D-5L preferences. Thus, as expected, the preferred model and the performance of these preferred models in terms of goodness-of-fit were quite different. For instance, the preferred model for the new English value set produced r^2, MAE and RMSE values of 0.342, 0.111 and 0.150, respectively, for DASS-21 compared with 0.332, 0.155 and 0.206 in the

previous study.[4] Similarly, the preferred model for the K-10 produced an r^2, MAE and RMSE of 0.337, 0.110 and 0.151, respectively, compared with 0.361, 0.150 and 0.201 in the previous study, indicating better predictive performance in our study. These differences in goodness-of-fit may, in part, be because of differences in the scale of the target instrument and the regression method applied. Third, we have shown that mapping functions will differ across countries depending on cross-cultural diversity in the preferences on which EQ-5D-5L value sets are based. In addition, different covariates have been used in the two studies. The previous study included country dummies and gender, whereas our study has considered respondents' age and gender alone.

A recent review of mapping studies found that the goodness-of-fit measured by r^2 ranges from 0.17 to 0.71, with most studies reporting an r^2 between 0.4 and 0.5.[26] A study by Lindkvist and Feldman[35] assessed mapping a mental health-specific outcome measure (12-item General Health Questionnaire) onto EQ-5D-3L with the UK and Swedish value sets. They reported an r^2 and RMSE of 0.18 and 0.20 for the UK value set, and 0.24 and 0.07 for the Swedish value set, respectively, when the 12-item General Health Questionnaire alone was used as a predictor. Another study by Brazier et al[36] mapped the Hospital Anxiety and Depression Scale onto EQ-5D-3L in two different samples. They reported an r^2 of 0.24 and RMSE of 0.227 in the first sample, and an r^2 of 0.19 and RMSE of 0.188 in the second sample. The mapping algorithm produced in our study showed better performance, although they differ in terms of methodological approach and predictor variables used.

Mapping algorithms generally suffer from overprediction of utility values for respondents in poor health and underprediction for respondents in better health.[26] This was also the case in our study (see supplementary Fig. 1). A possible reason for this may, in part, be a lack of conceptual overlap between the source instruments and EQ-5D-5L. For instance, as revealed by the EFA, only the anxiety/depression dimension of the EQ-5D-5L has been mainly loaded onto one of the same factors that the disease-specific outcomes were designed to measure. Another plausible reason would be the strong decrements of preference weights of the EQ-5D-5L at a severe health state, i.e. when moving from level 3 to level 4.[37] This study has explored the mapping algorithms for different value sets of EQ-5D-5L against depression scales. Because different EQ-5D-5L value sets produce different utility scores, especially at the lower end, the country-specific mapping algorithm should be a better option to reflect the preference from a particular country. Furthermore, this is the first study to assess the predictive accuracy of different EQ-5D-5L value sets with the DASS-21 and K-10 instrument. Considering the multinational nature of the patient population used, our algorithms may have wider generalisability. However, as generalisability is a major issue for mapping studies, it should be tested how these models perform in different patient populations.

This study has some limitations. First, it is based on respondents who volunteered to participate, something that might lead to self-selection bias. Second, as the EFA results indicated, the conceptual overlap between the source and target instruments is limited. However, if the generic instrument covers important dimensions of the source instrument, it is feasible to conduct mapping studies.[1] Although the physical dimensions of EQ-5D-5L are less correlated with DASS-21 and K-10, results from the EFA revealed conceptual overlap with the depression scales. Furthermore, studies have shown that EQ-5D reflects the effect of common mental health conditions such as mild to moderate depression,[4,38] suggesting that mapping depression scales onto EQ-5D is plausible.

In conclusion, this study has developed a set of mapping algorithms to predict EQ-5D-5L utility values from the DASS-21 or the K-10. Thus, in the absence of generic health-related quality of life data, the preferred mapping model can adequately convert disease-specific scores onto a generic outcome metric such as QALYs, which facilitates economic evaluations of mental health interventions.

Thor Gamst-Klaussen, MA, PhD Fellow; **Admassu N. Lamu**, PhD, Research Fellow, Department of Community Medicine, University of Tromsø, Norway; **Gang Chen**, PhD, Senior Research Fellow, Centre for Health Economics, Monash University, Australia; **Jan Abel Olsen**, PhD, Professor, Department of Community Medicine, University of Tromsø, Norway and Centre for Health Economics, Monash University, Australia

Correspondence: Thor Gamst-Klaussen, MA, Department of Community Medicine, PO Box 6050, University of Tromsø, 9037 Tromsø, Norway. Email: thor.klaussen@uit.no

Funding

Supplementary material

Supplementary material is available online at https://doi.org/10.1192/bjo.2018.21

References

1 Brazier J, Ratcliffe J, Salamon JTA. *Measuring and valuing health benefits for economic evaluation.* Oxford University Press, 2016.

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

3 World Health Organization. *Mental Disorders*. World Health Organization, 2017 (http://www.who.int/mediacentre/factsheets/fs396/en/).

4 Lovibond SH, Lovibond PF. *Manual for the Depression Anxiety Stress Scales*. Psychology Foundation, 1995.

5 Kessler RC, Barker PR, Colpe LJ, Epstein JF, Gfroerer JC, Hiripi E, et al. Screening for serious mental illness in the general population. *Arch Gen Psychiatry* 2003; **60**(2): 184–9.

6 Mihalopoulos C, Chen G, Iezzi A, Khan MA, Richardson J. Assessing outcomes for cost-utility analysis in depression: comparison of five multi-attribute utility instruments with two depression-specific outcome measures. *Br J Psychiatry* 2014; **205**(5): 390–7.

7 Wisloff T, Hagen G, Hamidi V, Movik E, Klemp M, Olsen JA. Estimating QALY gains in applied studies: a review of cost-utility analyses published in 2010. *Pharmacoeconomics* 2014; **32**(4): 367–75.

8 National Institute for Health and Care Excellence. *Guide to the Methods of Technology Appraisal* 2013. National Institute for Health and Care Excellence, 2013 (https://www.nice.org.uk/process/pmg9/).

9 van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3 L value sets. *Value Health* 2012; **15**: 708–15.

10 Dolan P. Modeling valuations for EuroQol health states. *Med Care* 1997; **35**(11): 1095–108.

11 Augustovski F, Rey-Ares L, Irazola V, Garay OU, Gianneo O, Fernandez G, et al. An EQ-5D-5L value set based on Uruguayan population preferences. *Qual Life Res* 2015; **25**: 323–33.

12 Kim S-H, Ahn J, Ock M, Shin S, Park J, Luo N, et al. The EQ-5D-5L valuation study in Korea. *Qual Life Res* 2016; **25**(7): 1845–52.

13 Luo N, Liu G, Li M, Guan H, Jin X, Rand-Hendriksen K. Estimating an EQ-5D-5L value set for China. *Value Health* 2017; **20**: 662–9.

14 Ramos-Goni JM, Pinto-Prades JL, Oppe M, Cabases JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. *Med Care* 2017; **55**: e51–8.

15 Versteegh MM, Vermeulen KM, Evers SMAA, de Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. *Value Health* 2016; **19**: 343–52.

16 Xie F, Pullenayegum E, Gaebel K, Bansback N, Bryan S, Ohinmaa A, et al. A time trade-off-derived value set of the EQ-5D-5L for Canada. *Med Care* 2016; **54**: 98–105.

17 Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. *Health Econ* 2018; **27**: 7–22.

18 Shiroiwa T, Ikeda S, Noto S, Igarashi A, Fukuda T, Saito S, et al. Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. *Value Health* 2016; **19**(5): 648–54.

19 Zhao Y, Li SP, Liu L, Zhang JL, Chen G. Does the choice of tariff matter?: A comparison of EQ-5D-5L utility scores using Chinese, UK, and Japanese tariffs on patients with psoriasis vulgaris in Central South China. *Medicine (Baltimore)* 2017; **96**(34): e7840.

20 Petrou S, Rivero-Arias O, Dakin H, Longworth L, Oppe M, Froud R, et al. Preferred reporting items for studies mapping onto preference-based outcome measures: the MAPS statement. *Pharmacoeconomics* 2015; **33**(10): 985–91.

21 Richardson J, Iezzi A, Maxwell A. *Cross-national Comparison of Twelve Quality of Life Instruments: MIC Paper 1 Background, Questions, Instruments. Research Paper 76*. Centre for Health Economics, Monash University, 2012 (http://www.buseco.monash.edu.au/centres/che/pubs/researchpaper76.pdf).

22 Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 2011; **20**(10): 1727–36.

23 Russell DW. In search of underlying dimensions: the use (and abuse) of factor analysis in personality and social psychology bulletin. *Pers Soc Psychol Bull* 2002; **28**(12): 1629–46.

24 Antony MM, Bieling PJ, Cox BJ, Enns MW, Swinson RP. Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychol Assess* 1998; **10**(2): 176–81.

25 Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. *Psychol Methods* 1999; **4**(3): 272–99.

26 Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *Eur J Health Econ* 2010; **11**: 215–25.

27 Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta distributed dependent variables. *Psychol Methods* 2006; **11**: 54–71.

28 Khan I, Morris S, Pashayan N, Matata B, Bashir Z, Maguirre J. Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patients. *Health Qual Life Outcomes* 2016; **14**: 60.

29 Papke LE, Wooldridge JM. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *J Appl Econom* 1996; **11**(6): 619–32.

30 Susanti Y, Sri Sulistijowai H, Pratiwi H, Liana T. M estimation, S estimation, and MM estimation in robust regression. *Int J Pure Appl Mathem* 2014; **91**(3): 349–60.

31 Ayinde K, Lukman AF, Arowolo O. Robust regression diagnostics of influential observations in linear regression model. *Open J Stat* 2015; **5**(4): 272–83.

32 Powell JL. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*. 1984; **25**(3): 303–25.

33 Versteegh MM, Leunis A, Luime JJ, Boggild M, Uyl-de Groot CA, Stolk EA. Mapping QLQ-C30, HAQ, and MSIS-29 on EQ-5D. *Med Decis Making* 2012; **32**: 554–68.

34 Sullivan PW, Ghushchyan V. Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. *Med Decis Making* 2006; **26**(4): 401–9.

35 Lindkvist M, Feldman I. Assessing outcomes for cost-utility analysis in mental health interventions: mapping mental health specific outcome measure GHQ-12 onto EQ-5D-3L. *Health Qual Life Outcomes* 2016; **14**(1): 134.

36 Brazier J, Connell J, Papaioannou D, Mukuria M, Mulhern B, Peasgood T, 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* 2014; **18**(34).

37 Olsen JA, Lamu A, Cairns J. In search of a common currency: a comparison of seven EQ-5D-5L value sets. *Health Econ* 2018; **27**(1): 39–49.

38 Brazier J. Is the EQ-5D fit for purpose in mental health? *Br J Psychiatry* 2010; **197**(5): 348–9.