Eur J Health Econ (2015) 16:847–855
DOI 10.1007/s10198-014-0633-1

ORIGINAL PAPER

Between-country heterogeneity in EQ-5D-3L scoring algorithms: how much is due to differences in health state selection?

Eleanor M. Pullenayegum · Kuhan Perampaladas · Kathryn Gaebel · Brett Doble · Feng Xie

Received: 2 July 2013 / Accepted: 29 August 2014 / Published online: 25 September 2014

Abstract

Background EQ-5D-3L scoring algorithms vary amongst countries, not only in the values of regression coefficients but also in the independent variables included in the regression model (hereafter referred to as model specification). It is unclear how much of this variation is due to differences in health state selection, the relative frequencies with which health states were valued, and model diagnostics, rather than to genuine differences in population preferences.

Methods Using aggregate data from a recent review, we noted all model specifications that were used. For each country the country's own model was re-fitted, as were all other model specifications. This was done twice: once using all valued health states for each country, and again using a common set of 17 health states for all countries. Goodness of fit was assessed using the following model diagnostics: mean absolute error (MAE), mean squared error (MSE) and rho (the Pearson correlation coefficient between predicted and observed mean utilities), both with and without leave-one-out cross-validation.

Results Thirteen countries contributed data. Even when using a common set of health states, the preferred model varied across countries. However, choice of health states did impact the preferred model specification: when using cross-validation, the preferred specification changed in five of ten countries when moving from 17 health states to all valued health states. The relative frequency with which health states were valued had little impact on the preferred model.

Conclusions Variation in choices of health states to value is responsible for some, but not all, of the observed heterogeneity in model specification. Relative frequency of health state valuation and choice of model diagnostic has a limited impact on model preference, however, use of cross-validation has a substantial impact. The use of cross-validation, implemented through omitting health states rather than respondents, is recommended as one approach to assessing model fit.

Keywords Health utility · EQ-5D · Scoring algorithm · Heterogeneity

JEL Classification I100 · C190

Introduction

Quality-adjusted life-years (QALYs) incorporate information on both morbidity and mortality, and are the preferred outcome in economic evaluations for both the National Institute for Clinical Excellence (NICE) in the UK and for the Canadian Agency for Drugs and Technology in Health [1, 2]. Health utilities are the quality weights used to calculate QALYs, and are typically measured using standardised questionnaires, of which the EQ-5D is very widely

Electronic supplementary material The online version of this article (doi:10.1007/s10198-014-0633-1) contains supplementary material, which is available to authorized users.

E. M. Pullenayegum (✉)
Child Health Evaluative Sciences, Hospital for Sick Children,
Toronto, ON, Canada
e-mail: eleanor.pullenayegum@sickkids.ca

K. Perampaladas · K. Gaebel · F. Xie
Department of Clinical Epidemiology and Biostatistics,
McMaster University, Hamilton, ON, Canada

B. Doble
Faculty of Business and Economics, Centre for Health
Economics, Monash University, Clayton, VIC, Australia

used [3]. The EQ-5D-3L questionnaire contains questions on five dimensions, namely mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. For each of these dimensions, the respondent is asked to indicate their level of difficulty, with response options being "no problems", "some problems", or "extreme problems". EQ-5D-3L responses must be converted to health utilities via a scoring algorithm. The scoring algorithm is developed in a valuation study in which respondents from the general population provide preferences for a subset of health states, and regression models are used to describe the mean utilities as a function of the health states.

The first scoring algorithm was developed in the UK by Dolan [4] in the Measurement and Valuation of Health (MVH) study. Since then, different scoring algorithms have been developed for different countries. Many of these studies have reported significant differences between their own algorithm and the original UK algorithm [5–7]. They differ not just in the values of regression coefficients, but also in the independent variables included in the regression model (hereafter referred to as model specification). Given that health preferences may depend on respondent age, gender, income and self-reported health status [8], and that these population characteristics vary by country, it is not surprising that regression coefficients vary by country. It is less clear that this would also lead to variation in the model specification. The MVH study used an N3 term [4], whilst the USA study used a D1 model [9], and the South Korean algorithm used a main effects model with a log transformation [10]. It is not known how much of this variation in model specification is due to genuine differences in health preferences amongst countries, as opposed to differences in health state selection, the relative frequencies with which states were valued, or the model diagnostics used to select the preferred model.

Whilst some studies used the MVH protocol [4], many use variants. Notably, the subset of states included and the relative frequencies with which states were valued vary among studies. For example, studies in the UK, Spain, USA, South Korea and Chile all used 42 health states [4, 5, 9–11], whilst studies in Japan and the Netherlands used a modified protocol [12] assessing 17 health states. Simulation evidence from Lamers [13] was used to justify a reduction in the total number of health states valued in the Dutch valuation study; the simulation was based on MVH data and thus assumed that the model specification in the Netherlands would be similar to that in the UK. A further simulation study by Chuang and Kind [14] based on a main effects model suggests that when there are restrictions on study size, it is preferable to include fewer health states, up to a minimum of 31. A recent simulation study by Viney [15] suggests that it is reasonable to expect that studies that incorporate more health states are more likely to detect interactions, and this was used to justify the use of 198 health states in the Australian valuation study. There are thus conflicting recommendations in the literature on the number of health states that should be valued.

Furthermore, the relative frequencies with which states were valued vary; in any given valuation study it is common to find some states which are valued more often than others. For example, the state 33333 was valued roughly 4 times as often as the state 12111 in Denmark [16], compared to 10 times as often in Canada [17], 3 times as often in the USA [9], twice as often in Poland [18] and roughly the same number of times in Japan [7]. Further, studies have differed in their model diagnostics. For example, some use the mean absolute error (MAE), whilst others use an R^2 or an adjusted R^2. Recent comparisons between countries have not been able to disentangle genuine differences in cultural preferences from differences in methodology [19].

The aim of this study was to assess the extent to which health state selection, frequency of health states valued, and model diagnostics have contributed to between-country heterogeneity in model specification among EQ-5D-3L scoring algorithms.

Methods

Systematic review

This study is based on a recent systematic review by Xie et al. [20]. The inclusion criteria for the original review were that studies should have (1) used elicitation techniques to obtain preferences for at least a subset of the EQ-5D-3L health states, and (2) explicitly indicated the preferred scoring algorithm to predict utilities for all EQ-5D-3L health states. For the present review, we included only those studies that used time trade-off (TTO) elicitation techniques and that reported mean observed utilities for each state that was valued. For each state included in each valuation study, data were extracted on the mean utility assigned to that state, the corresponding standard deviation, and the number of respondents valuing that state (see Supplementary Table 1). The final scoring algorithm was recorded.

Statistical analysis

Each of the final country-specific model specifications was re-fitted, both for the country on which it was originally derived and for all other countries. We had access to aggregate data only, that is, for each country we had only the mean observed utility for each valued health state. This is, however, sufficient to obtain unbiased estimates of regression coefficients using ordinary least squares (OLS). We regressed observed mean utilities for the health states onto characteristics of the health states. We began by restricting attention to those studies that included the set of 17 health states indicated by the modified MVH protocol [12]. Models were fitted using data from just these 17 health states. Thus, in this analysis all included countries used the same set of health states and each state received equal weight (using OLS on the aggregate data achieves equal weighting for each health state). The mean absolute error (MAE, i.e. the mean across-health states of the absolute difference between observed and predicted mean utilities), mean squared error (MSE) and rho (i.e. the correlation between observed and predicted utilities) for the country-specific model were compared to those for the other countries' model specifications. This was done both with and without cross-validation. To implement cross-validation, each of the 17 health states was omitted from the model in turn, and the fit from the remaining 16 health states was used to calculate the diagnostics for the omitted health state. Thus, the diagnostics represent out-of-model prediction errors, which we shall refer to as "leave-a-state-out cross-validation".

This analysis was repeated using all included countries and all valued health states, again using each reported model specification for each country. This allowed us to assess whether preference for the country's own model specification changes when states beyond the set of 17 are included. Finally, to assess whether preference for the country's own model changes when some states are represented more often than others, the analysis using all health states was repeated using weighted least squares (WLS) rather than OLS, with weights proportional to the number of times each state was valued. This reflects the weighting that would be given to each state in the original respondent-level analysis. Indeed, WLS on the aggregate data with weights equal to the number of respondents valuing each state is identical to OLS on the individual-level data.

We also computed adjusted R^2 statistics for models fitted with WLS, using both the restricted set of 17 health states and all valued health states. Since the conceptual basis for the adjusted R^2 assumes that the model has an intercept, this was only done for those models that included an intercept.

Finally, we checked all fitted models for logical consistency. We identified all pairs of health states in which one state was dominant, and compared the model's predicted mean utility for the dominant health state to that for the dominated health state. A model was denoted as yielding a logically inconsistent value set if there was at least one health state pair with a dominant health state whose utility was lower than the utility for the dominated health state.

Results

This review contains data from 13 countries [4, 5, 7, 9, 11, 13, 16–18, 21–24], see Table 1. Data from Kind [25] was excluded because it was a subgroup analysis (England and Wales vs Scotland) using the data from Dolan [4]. Similarly, results from Shaw et al. [26] were excluded as this used the same data as in the original study [9], but using median regression in the place of mean regression, and results from Zarate [27] were excluded as this split the Shaw [9] results into subgroups (Hispanics vs others). Since the Taiwanese study of Chang [28] used a minimal set of 13 health states and was designed to fit the main effects model only, this was also omitted from the analysis. The South Korean study [10] had to be excluded as this used a log transformation on the individual level utilities, which we were not able to replicate using aggregate data.

The 13 countries used a total of seven model specifications (Table 1). Five countries used main effects models [7, 16–18, 24], and five countries included an N3 term, with Germany using a reduced N3 model that omitted some of the main effects, and France omitting the intercept. Three countries used models that included interaction terms other than the N3 term [9, 11, 21].

Model preference using cross-validation

The choice of preferred model was similar between the MAE, MSE and rho, with model preference being the same across diagnostics in nine of the 13 countries (see Table 2). The MAE and MSE had identical preferred models in all but four cases, exceptions being Canada and the USA when using all health states and fitting with OLS, and Poland and the USA when using all health states and fitting with WLS. Notable differences in preference between rho and the MAE or MSE diagnostics were a shift from preference for the German model in the Netherlands when using the MAE or MSE to the French model when using rho, and a shift in the USA away from the German and Chilean models when using the MAE or MSE to the Argentinean model when using rho. Given the similarities in model preferences between diagnostics, in what follows we use the MAE in assessing the impact of using differing health states and weightings on model preference.

There were ten studies that reported mean utilities for the set of 17 health states [4, 5, 7, 9, 13, 17, 18, 21–23]. Table 2 gives the preferred models, whilst Table 3 gives the MAE for each model considered. On re-fitting both the country's own model specification and the other specifications using the 17 states, the country's own model was preferred in four countries (Chile, France, Japan and Poland). Five countries (Argentina, Canada, the Netherlands, the UK and the USA) showed a preference for the

Table 1 Studies included in the analysis, with their own functional forms

Country	Number of states valued	Sample size analysed	Own model	Intercept	Ten main effects	N3	Other terms
Argentina	22	611	Argentina		✔		O2, O3, Z2, Z3, $C2^2$, $C3^2$
Canada	48	1,145	Main effects	✔	✔		
Chile	42	1,967	Chile	✔	✔		$C3^2$, X5
Denmark	46	1,332	Main effects	✔	✔		
France	25	443	France		✔	✔	
Germany	36	334	Germany	✔		✔	MO2, MO3, SC2, SC3, PD2, PD3, AD3
Japan	17	543	Main effects	✔	✔		
Netherlands	17	298	N3	✔	✔	✔	
Poland	44	305	Main effects	✔	✔		
Spain	43	975	N3	✔	✔	✔	
UK	42	2,997	N3	✔	✔	✔	
USA	42	3,773	D1		✔		D1, $I2^2$, I3, $I3^2$
Zimbabwe	38	2,384	Main effects	✔	✔		

MO2 dummy variable for mobility at level 2, *MO3* dummy variable for mobility at level 3, *SC2* dummy variable for self-care at level 2, *SC3* dummy variable for self-care at level 3, *PD2* dummy variable for pain/discomfort at level 2, *PD3* dummy variable for pain/discomfort at level 3, *AD3* dummy variable for anxiety/depression at level 3, *N3* term 1 if any dimension is at level 3, 0 otherwise, *D1* additional number of dimensions at either level 2 or level 3, *I2* number of dimensions at level 2 beyond the first, *I3* number of dimensions at level 3 beyond the first, *O2* 1 if all dimensions at level 1 and level 2, 0 otherwise, *O3* 1 if all dimensions at level 1 and level 3, 0 otherwise, *Z2* 1 if at least one dimension at level 2 and one dimension at level 3, 0 otherwise, *Z3* number of dimensions at level 2 given at least one dimension at level 3, *C2* number of dimensions at level 2, *C3* number of dimensions at level 3, *X5* 1 if all 5 dimensions at either level 2 or level 3

German reduced N3 model, whilst Spain preferred the Chilean model. Thus, among the ten countries including the set of 17 health states, heterogeneity in model specification was reduced from six specifications to four when a common set of health states was used.

When all valued health states were used, we were able to use a further three studies [16, 23, 24]. On re-fitting the country's own model and the other models, there was a preference for the country's own model in four of the 13 countries (Chile, Germany, Japan, and Poland), see Tables 2 and 3. The Chilean model was preferred in three countries besides Chile (Denmark, the UK and the USA), the Argentinian model was preferred in three countries (France, Spain and Zimbabwe), the N3 model was preferred in one country (Canada), and the German reduced N3 model was preferred in two countries besides Germany (Argentina and the Netherlands). There were five preferred model specifications across the 13 studies.

When all health states were used and weighted least squares was used to fit the models, we had to exclude four studies [5, 16, 21, 24] as they did not indicate the number of valuations per respondent (this was the weight used in the weighted least squares models). Three of the nine countries favoured their own models (Chile, Germany and Japan). The Netherlands continued to favour the German reduced N3 specification, whilst France favoured the Argentinian specification. There was a very marginal preference for the D1 model in Poland (see Tables 2, 3). Canada, the UK and the USA favoured the Chilean model. The preferred model specification was similar to that using all health states and ordinary least squares, indicating that the number of valuations per health state had little impact on heterogeneity in model specification.

Compared to the countries' own models, there were substantial improvements in fit on using an alternative specification in several countries. For example, in Argentina the proportion of health states with predictions differing from observed values by more than 0.05 decreased from 77 % on using the Argentinian specification to 68 % when using the German specification (using all health states and OLS), see Table 4. For the Netherlands, 76 % of health states had predicted values differing from the observed by more than 0.05 on using the N3 specification that was used for the Dutch value set, compared to 65 % on using the German specification.

Model preference without cross-validation

In the absence of any cross-validation, there was an overwhelming preference for the Argentinian, Chilean and N3 models. When there was no penalty for the number of parameters in the model, the Argentinian model was preferred regardless of whether the MAE, MSE or rho was used, regardless of whether all health states or just 17 health states were used and regardless of whether OLS or WLS were used; the only exception to this was that the

Table 2 Preferred model using mean absolute error (MAE) and leave-a-state-out cross-validation, mean squared error (MSE) and leave-a-state-out cross-validation, and adjusted R^2 with no cross-validation

Country	Own model	Preferred model using MAE			Preferred model using MSE			Preferred model using rho			Preferred model using adjusted R^2	
		17	All-OLS	All-WLS	17	All-OLS	All-WLS	17	All-OLS	All-WLS	17	All
Argentina	A	G	G	N/E	G	G	N/E	G	G	N/E	N/E	N/E
Canada	ME	G	N3	Ch	G	ME	Ch	G	Ch	Ch	Ch (N3)	Ch
Chile	Ch	Ch (N3)	Ch	Ch	Ch (N3)	Ch	Ch	Ch (F)	Ch	Ch	Ch (N3)	Ch
Denmark	ME	N/E	Ch (N3)	N/E	N/E	Ch (G)	N/E	N/E	Ch (G)	N/E	N/E	N/E
France	F	F	A (F)	A (F)	F	A (F)	A (F)	F	A (F)	A (F)	N3	N3
Germany	G	N/E	G	G	N/E	G	G	N/E	G	G	N/E	N3 (G)
Japan	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	N3	N3
Netherlands	N3	G	G	G	G	G	G	F	F	F	N3	N3
Poland	ME	ME	ME	D1	ME	ME	ME	ME	ME	ME	Ch (N3)	Ch
Spain	N3	Ch (F)	A (D1)	N/E	Ch (F)	A (D1)	N/E	Ch (F)	A (D1)	N/E	N/E	N/E
UK	N3	G	Ch (D1)	Ch (D1)	G	Ch (N3)	Ch (D1)	G	Ch (D1)	Ch (D1)	N3	Ch (N3)
USA	D1	G	Ch (N3)	Ch (N3)	G	A (N3)	A (N3)	A (G)	Ch (N3)	D1 (N3)	N3	Ch (N3)
Zimbabwe	ME	N/E	A (ME)	N/E	N/E	A (ME)	N/E	N/E	A (ME)	N/E	N/E	N/E

Note that without cross-validation, model preference using the MAE, MSE or rho was universally for the Argentinian model, with the exception of the Chilean model for Chile on using the MAE. In cases where the preferred model was logically inconsistent, the preferred model among those models that are logically consistent is indicated in parentheses

* When calculating the adjusted R^2, the following models could not be considered: Argentinian, French, and D1 models (no intercept). *N/E* not estimable. In the case of models based on the 17 health states, models were not estimable in those countries that did not collect data on all 17 health states. In the case of weighted least squares models, models were not estimable in those countries that did not report the number of valuations per health state, as this number was used as the weight

Chilean model was preferred in Chile when the MAE was used on all health states using OLS.

When the number of parameters in the model was accounted for using the adjusted R^2, countries that did not report the number of valuations per health state had to be excluded [5, 16, 21, 24] and, moreover, models that omitted an intercept (France, D1, Argentina) could not be considered. Here the preference was for Chilean and N3 models, with the N3 model more often preferred when using just 17 health states.

Logical consistency

As can be seen from Table 3, all model specifications led to logically inconsistent results for at least one country. The Argentinian model yielded logically inconsistent value sets in the majority of cases, exceptions being Chile and Poland when all health states were used. The Chilean model also yielded logically inconsistent value sets in a number of cases, particularly when just 17 health states were used to fit the model. It was common that one of the Argentinian or Chilean specifications was preferred, using the MAE, MSE or rho, but also yielded a logically inconsistent value set. As can be seen from Table 2, restricting attention to those specifications that yield logically consistent value sets does not change the finding that model diagnostic (MAE, MSE, rho) and weighting have little impact on model choice. This also held when model diagnostics were calculated without cross-validation.

Discussion

This analysis has investigated the impact of health state selection, frequency of health states valued, and model diagnostics on heterogeneity in model specification amongst the existing EQ-5D-3L algorithms. In terms of model diagnostics, there was little difference in model preference between the MAE and MSE. However, use of the adjusted R^2 altered model preference. When cross-validation was adopted, the preferred model specification changed when moving from a common set of 17 health states to using all valued health states. Thus, health state selection has an impact on the preferred model specification. The relative frequencies with which states were valued contributed little to the heterogeneity in model specification.

Use of leave-a-state-out cross-validation resulted in heterogeneity in model specification, whereas omitting cross-validation led to homogeneity in specification. The reduction in heterogeneity on omitting cross-validation should not be taken as an endorsement of the practice,

Table 3 Mean absolute errors calculated by computing, for each health state, the absolute difference between the observed mean and the model prediction for each functional form, then taking the mean across health states

Country	Own	N3	Main effects	D1	Argentina	Chile	German	French
17 health states using Ordinary Least Squares								
Argentina	0.481	0.244	0.230	0.384	0.481	0.286	**0.138**	0.203
Canada	0.093	0.094	0.093	0.235	0.205	0.111	**0.087**	0.119
Chile	**0.056**	*0.101*	0.122	0.116	0.258	**0.056**	0.151	0.109
France	**0.097**	0.137	0.163	0.254	0.634	0.173	0.115	**0.097**
Japan	**0.044**	0.047	**0.044**	0.053	0.168	0.067	0.080	0.176
Netherlands	0.090	0.090	0.172	0.112	0.157	0.147	**0.083**	0.099
Poland	**0.097**	0.149	**0.097**	0.214	0.246	0.159	0.123	0.111
Spain	0.096	0.096	0.225	0.118	0.404	**0.081**	0.102	*0.091*
UK	*0.121*	*0.121*	0.220	0.127	0.201	0.167	**0.075**	0.126
US	0.148	0.118	0.158	0.148	0.039	0.163	**0.066**	0.137
All health states using Ordinary Least Squares								
Argentina	0.198	0.137	0.153	0.169	0.198	0.169	**0.115**	0.163
Canada	0.050	**0.049**	0.050	0.053	0.054	0.049	0.064	0.061
Chile	**0.024**	0.044	0.052	0.031	0.029	**0.024**	0.097	0.048
Denmark	0.072	*0.071*	0.072	0.076	0.080	**0.064**	0.074	0.074
France	*0.070*	0.073	0.088	0.092	**0.066**	0.083	0.088	*0.070*
Germany	**0.053**	0.060	0.107	0.059	0.062	0.061	**0.053**	0.057
Japan	**0.044**	0.047	**0.044**	0.053	0.168	0.067	0.080	0.176
Netherlands	0.090	0.090	0.172	0.112	0.157	0.147	**0.083**	0.099
Poland	**0.044**	0.046	**0.044**	0.046	0.051	0.046	0.092	0.046
Spain	0.056	0.056	0.083	*0.052*	**0.049**	0.051	0.078	0.058
UK	0.052	0.052	0.078	*0.051*	0.049	**0.047**	0.058	0.065
US	0.039	*0.044*	0.052	0.039	0.034	**0.034**	0.049	0.066
Zimbabwe	*0.051*	0.058	*0.051*	0.052	**0.046**	0.063	0.064	0.076
All health states using Weighted Least Squares								
Canada	0.058	0.053	0.058	0.053	0.054	**0.049**	0.065	0.063
Chile	**0.024**	0.051	0.067	0.031	0.028	**0.024**	0.096	0.054
France	0.071	0.073	0.090	0.092	**0.067**	0.082	0.087	*0.071*
Germany	**0.055**	0.062	0.129	0.060	0.063	0.064	**0.055**	0.059
Japan	**0.044**	0.047	**0.044**	0.053	0.168	0.067	0.080	0.176
Netherlands	0.090	0.090	0.172	0.112	0.157	0.147	**0.083**	0.099
Poland	0.045	0.047	0.045	**0.045**	0.050	0.046	0.092	0.047
UK	0.056	0.056	0.098	*0.051*	0.048	**0.045**	0.059	0.067
US	0.038	*0.046*	0.064	0.038	0.033	**0.031**	0.049	0.068

All predictions used leave-one-out cross-validation. For each country, the lowest MAE is printed in bold. Models which gave logically inconsistent value sets are shaded in grey, and in cases where the model with the lowest MAE was logically inconsistent, the model with the lowest MAE among those that are logically consistent is indicated in italics

however. There is a strong conceptual argument in favour of cross-validation: whilst the aim of a valuation study is to estimate utilities for each of the 243 health states described by the EQ-5D-3L, typically valuation studies have included at most 42 health states (fewer than 20 % of the total number of health states). Thus, the model is used predominantly to make out-of-state predictions, and so the predictive accuracy of these predictions should be assessed. Furthermore, in the absence of cross-validation, one should expect a model with more parameters to yield better MAEs, MSEs and rhos, even if the additional parameters do not reflect any genuine patterns in the data. This is a plausible explanation for the preference for the Argentinian model (consisting of 17 parameters) over all other models when using MSEs, MAEs and rhos and omitting cross-validation. Use of the adjusted R^2, which does include a penalty for the number of parameters, resulted in preferences for either the N3 or Chilean models. Part of the

Table 4 Percentages of health states with out-of-sample predictions differing from observed means by more than 0.1 and 0.05

Country	% of predictions differing from observed means by more than 0.1						% of predictions differing from observed means by more than 0.05					
	17 states, OLS		All states, OLS		All states, WLS		17 states, OLS		All states, OLS		All states, WLS	
	Preferred	Own	Preferred	Own	Preferred	Own	Preferred	Own	Preferred	Own	Preferred	Own
Argentina	53	100	54	68	N/E	N/E	71	100	68	77	N/E	N/E
Canada	41	35	19	15	17	21	65	65	40	38	38	44
Chile	12 (47)	12	0	0	0	0	35 (76)	35	10	10	10	10
Denmark	N/E	N/E	20 (20)	20	N/E	N/E	N/E	N/E	59 (54)	59	N/E	N/E
France	47	47	28 (28)	28	28 (28)	28	76	76	56 (60)	60	56 (60)	60
Germany	N/E	N/E	9	9	9	9	N/E	N/E	43	43	46	46
Japan	12	12	12	12	12	12	29	29	29	29	29	29
Netherlands	29	41	29	41	29	41	65	76	65	76	65	76
Poland	47	47	14	14	11	11	71	71	39	39	27	36
Spain	35 (47)	47	19 (17)	19	N/E	N/E	71 (71)	76	38 (36)	40	N/E	N/E
UK	29	53	7 (10)	10	7 (12)	12	65	94	36 (40)	40	33 (40)	43
USA	18	89	7 (5)	7	7 (10)	7	65	82	19 (38)	33	17 (45)	33
Zimbabwe	N/E	N/E	5 (11)	11	N/E	N/E	N/E	N/E	35 (38)	38	N/E	N/E

As model preferences were similar across model diagnostics when using cross-validation, only model preference using the MAE is considered here. In cases where the preferred model was logically inconsistent, the percentages for the preferred model among those models that were logically inconsistent are given in parentheses

N/E not estimable, *OLS* ordinary least squares, *WLS* weighted least squares

relative homogeneity on using the adjusted R^2 may have been due to the fact that the adjusted R^2 could not be computed for four of the seven models.

When just 17 health states were used, the preferred model was most often the German reduced N3 or main effects model, whilst when more health states were used, preference often switched to the more complex Argentinian or Chilean models. This is to be expected theoretically, and was demonstrated empirically by Viney [15]; the basic intuition behind the result is that one is better able to detect interaction terms when more health states are represented. The reason for the introduction of the 17-state protocol was due to perceived redundancies in the original 42 states, coupled with a desire to simplify the process of a valuation study [12]. This decision was, however, based on UK data. Subsequent simulation evidence from Lamers et al. [13] also used UK data when selecting health states, and in particular calculated mean absolute errors averaged over a mixture of within-sample and out-of-sample health states, considering only the N3 model.

An important qualification on preference for the Argentinian and Chilean specifications is that in many cases they yielded logically inconsistent value sets despite having the lowest MAE or MSE. Indeed, the published value set for Argentina is itself logically inconsistent for a number of pairs of health states (see, for example the pairs (33111, 33112) and (33321, 33322) in [21]). If a logically consistent value set is desired, we suggest that logical consistency be checked when using these specifications, especially when used in a valuation study that includes a limited number of health states.

The main limitation of this analysis was the use of aggregated data rather than respondent-level data. This meant that we were not able to test differences in regression diagnostics between algorithms for statistical significance within countries, as the aggregated data meant that it was not possible to account for the correlation that arises from each respondent valuing multiple health states. Furthermore, the use of aggregated data meant that it was not possible to consider fitting models using random effects. While we were able to consider cross-validation omitting health states, we were not able to consider cross-validation omitting respondents. For large valuation studies, however, it is more important to consider omitting states than it is to omit respondents. Cross-validation omitting respondents is typically implemented in such a way as to estimate the prediction errors at the individual level; cross-validation omitting health states, as done here, estimates the error in mean utility at the population level.

The use of aggregate data is not as severe a limitation as one might think. It is straightforward to show that OLS and WLS models provide unbiased estimates of regression coefficients even in the presence of within-subject correlation. Moreover, WLS on the aggregate data with weights equal to the number of respondents valuing each state is identical to OLS on the individual-level data. We used both

OLS and WLS in our analysis as doing so enabled us to isolate the effects of health state selection and frequency of valuation on heterogeneity in model specification. For example, since OLS on the aggregate data weights all health states equally, comparing OLS using all health states to WLS using all health states allowed us to examine the effect of varying frequency of health state valuation.

A further limitation was the exclusion of several studies such as the Australian valuation study [15] due to unavailability of key data (for example, mean and standard error of the TTO utilities attached to each valued health state). Whilst this limits the number of studies available for analysis, it is unlikely to have introduced bias. Finally, there are other differences in protocol, for example the population sampled, the strictness of the exclusion criteria and the method of transformation for states considered worse than dead. Only Shaw [9] used a transformation other than the original Dolan transformation for states valued as worse than dead, and so it is unlikely that the observed heterogeneity in model specification is due to differences in transformation. Most studies sampled the general population, exceptions being the Argentinian study, which used patients and family members, and the Polish study, which used visitors to inpatients. The exclusion criteria varied considerably. These and other differences may be responsible for some of the observed heterogeneity in model specification.

We found that the choice of health states to value is responsible for some, but not all, of the observed heterogeneity in model specification. Specifically, our results showed that even when the same states and weightings are used in each country, there is heterogeneity in model specification. This finding underscores the importance of a common valuation protocol for EQ-5D valuation studies as a means of reducing model heterogeneity amongst countries, but also suggests that heterogeneity in specification should still be expected despite the common protocol.

The finding that the country's own specification was often out-performed by alternative specifications has potentially important implications. There are at least two possible explanations for this finding. Firstly, some alternative specifications may not have been considered. This is not critical as no study claims to have found the best specification. Secondly, although several studies used cross-validation to assess model fit, in most cases this was done by omitting respondents. Only Dolan [4] used leave-a-state-out cross-validation. This is striking given that the ability to predict mean utilities for health states not included in the analysis is much more critical than the ability to predict utilities for individuals who are not included in the analysis. Furthermore, in cases where the country's own specification was out-performed by an alternative specification, the differences in predictions from the two models were large enough to be important for a significant proportion of the health states.

In conclusion, this analysis underlines the importance of health state selection when designing a valuation study, and suggests that cross-validation through the omission of states, rather than respondents, should be considered when assessing model fit.

Acknowledgments Eleanor Pullenayegum and Feng Xie are funded by the Canadian Institutes of Health Research New Investigator Salary Awards. Feng Xie is also funded by the Ontario Health Quality Council, and Eleanor Pullenayegum is funded by the Natural Sciences and Engineering Research Council.

References

1. National Institute for Health and Care Excellence: Guide to the methods of technology appraisal. National Institute for Health and Care Excellence, London, UK. http://publications.nice.org.uk/pmg9 (2013)
2. Canadian Agency for Drugs and Technology in Health: Guidelines for economic evaluation of health technologies, 3rd edn. Canadian Agency for Drugs and Technology in Health, Ottawa, ON, Canada (2006)
3. EuroQol–a new facility for the measurement of health-related quality of life.: The EuroQol group. Health Policy. 16, 199–208 (1990)
4. Dolan, P.: Modeling valuations for EuroQol health states. Med. Care **35**, 1095–1108 (1997)
5. Badia, X., Roset, M., Herdman, M., Kind, P.: A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med. Decis. Making **21**, 7–16 (2001)
6. Johnson, J.A., Luo, N., Shaw, J.W., Kind, P., Coons, S.J.: Valuations of EQ-5D health states: are the United States and United Kingdom different? Med. Care **43**, 221–228 (2005)
7. Tsuchiya, A., Ikeda, S., Ikegami, N., et al.: Estimating an EQ-5D population value set: the case of Japan. Health Econ. **11**, 341–353 (2002)
8. Shaw, J.W., Johnson, J.A., Chen, S., Levin, J.R., Coons, S.J.: Racial/ethnic differences in preferences for the EQ-5D health states: results from the USA valuation study. J. Clin. Epidemiol. **60**, 479–490 (2007)
9. Shaw, J.W., Johnson, J.A., Coons, S.J.: USA valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med. Care **43**, 203–220 (2005)
10. Jo, M.W., Yun, S.C., Lee, S.I.: Estimating quality weights for EQ-5D health states with the time trade-off method in South Korea. Value Health **11**, 1186–1189 (2008)
11. Zarate, V., Kind, P., Valenzuela, P., Vignau, A., Olivares-Tirado, P., Munoz, A.: Social valuation of EQ-5D health states: the Chilean case. Value Health **14**, 1135–1141 (2011)
12. Macaran S, Kind P.: Valuing EQ-5D health states using a modified MVH protocol. Prelim. Results 16, 205–39 (1999)
13. Lamers, L.M., McDonnell, J., Stalmeier, P.F., Krabbe, P.F., Busschbach, J.J.: The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. **15**, 1121–1132 (2006)
14. Chuang, L.H., Kind, P.: The effect of health state selection on the valuation of EQ-5D. Med. Decis. Making **31**, 186–194 (2011)
15. Viney, R., Norman, R., King, M.T., et al.: Time trade-off derived EQ-5D weights for Australia. Value Health **14**, 928–936 (2011)

16. Wittrup-Jensen, K.U., Lauridsen, J., Gudex, C., Pedersen, K.M.: Generation of a Danish TTO value set for EQ-5D health states. Scand. J. Public Health **37**, 459–466 (2009)
17. Bansback, N., Tsuchiya, A., Brazier, J., Anis, A.: Canadian valuation of EQ-5D health states: preliminary value set and considerations for future valuation studies. PLoS One **7**, e31115 (2012)
18. Golicki, D., Jakubczyk, M., Niewada, M., Wrona, W., Busschbach, J.J.: Valuation of EQ-5D health states in poland: first TTO-based social value set in central and eastern Europe. Value Health **13**, 289–297 (2010)
19. Norman, R., Cronin, P., Viney, R., King, M., Street, D., Ratcliffe, J.: International comparisons in valuing EQ-5D health states: a review and analysis. Value Health **12**, 1194–1200 (2009)
20. Xie F, Gaebel K, Perampaladas K, Doble B, Pullenayegum EM.: Comparing EQ-5D valuation studies: a systematic review and methodological reporting checklist. Med. Decis. Making 22 Mar 2013 (epub ahead of print)
21. Augustovski, F.A., Irazola, V.E., Velazquez, A.P., Gibbons, L., Craig, B.M.: Argentine valuation of the EQ-5D health states. Value Health **12**, 587–596 (2009)
22. Chevalier J, de Pouvourville G.: Valuing EQ-5D using time trade-off in France. Eur. J. Health. Econ. (2011)
23. Greiner, W., Claes, C., Busschbach, J.J., von der Schulenburg, J.M.: Validating the EQ-5D with time trade-off for the German population. Eur. J. Health. Econ. **6**, 124–130 (2005)
24. Jelsma, J., Hansen, K., De Weerdt, W., De Cock, P., Kind, P.: How do Zimbabweans value health states? Popul. Health Metr. **1**, 11 (2003)
25. Kind, P.: Valuing health benefits using the EQ-5D: the West Lothian question. In: Stavem, K. (ed.) 22nd Plenary meeting of the EuroQol Group, Oslo, Norway, pp. 55–77 (2005)
26. Shaw, J.W., Pickard, A.S., Yu, S., et al.: A median model for predicting United States population-based EQ-5D health state preferences. Value Health **13**, 278–288 (2010)
27. Zarate, V., Kind, P., Chuang, L.H.: Hispanic valuation of the EQ-5D health states: a social value set for Latin Americans. Value Health **11**, 1170–1177 (2008)
28. Chang, T.J., Tarn, Y.H., Hsieh, C.L., Liou, W.S., Shaw, J.W., Chiou, X.G.: Taiwanese version of the EQ-5D: validation in a representative sample of the Taiwanese population. J. Formos. Med. Assoc. **106**, 1023–1031 (2007)