The University Of Sheffield.

School of Health And Related Research

Health Economics & Decision Science (HEDS) Discussion Paper Series

Copula-based modelling of self-reported health states
An application to the use of EQ-5D-3L and EQ-5D-5L in evaluating drug therapies for rheumatic disease

Authors: Mónica Hernández Alava, Stephen Pudney

Corresponding author: Stephen Pudney

ISER, University of Essex,
Wivenhoe Park,
Colchester,
CO4 3SQ,
tel. +44(0)1206-873789;
email: spudney@essex.ac.uk

No. 16.06

Copula-based modelling of self-reported health states An application to the use of EQ-5D-3L and EQ-5D-5L in evaluating drug therapies for rheumatic disease

Mónica Hernández-Alava
School of Health and Related Research, University of Sheffield
Stephen Pudney
Institute for Social and Economic Research, University of Essex

This version February 16, 2016

Abstract

EQ-5D is used in cost-effectiveness studies underlying many important health policy decisions. It comprises a survey instrument generating a description of health states across five domains, and a system of utility values for each state. The original 3-level version of EQ-5D is being replaced with a more sensitive 5-level version but there little is known about the consequences of this change. We develop a multi-equation ordinal response model incorporating a copula specification with normal mixture marginals to analyse the joint responses to EQ-5D-3L and EQ-5D-5L in a survey of people affected by rheumatoid disease, and use it to generate mappings between the 3-level and 5-level descriptive systems. We find significant conflicts between the two, which would imply the reversal of an important conclusion in a real-world evaluation of drug therapies.

Keywords: EQ-5D, ordinal response, copula, mixture models, rheumatoid arthritis, mapping, economic evaluation

JEL codes: C35, C83, D61, H51, I10

Contact: Steve Pudney, ISER, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK; tel. +44(0)1206-873789; email: spudney@essex.ac.uk

This work was supported by the Medical Research Council under grant MR/L022575/1. Pudney acknowledges further ESRC funding through the UK Centre for Longitudinal Studies and the Centre for Micro-Social Change (grants RES-586-47-0002 and RES-518-28-5001). The authors wish to thank Kaleb Michaud (University of Nebraska Medical Center and National Data Bank for Rheumatic Diseases) and Frederick Wolfe (National Data Bank for Rheumatic Diseases) for providing the data, and Anastasios Panagiotelis and Allan Wailoo for helpful discussion. The views expressed in this article, and any errors or omissions, are those of the authors only.

1 Introduction: EQ-5D-3L and EQ-5D-5L

The quality-adjusted life year (QALY) is one of the most widely used health benefit measures in economic evaluations of interventions, services or programmes designed to improve health. The QALY allows health care decision makers to use a consistent approach across a broad range of disease areas, treatments, and patients. It reflects concerns for both quality and length of life and, in England, is the preferred outcome measure for the National Institute for Health and Care Excellence (NICE) in its appraisals of health interventions NICE (2014). Preference-based measures such as the EQ-5D underpin the calculation of QALYs by providing a simple descriptive profile and a single index value for health states.

EQ-5D measures patient health across five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The original version of EQ-5D, which has been used in a large number of cost-effectiveness evaluations, measures each domain on a scale with three severity levels (no problems, some or moderate problems, extreme problems). Up to $3^5 = 243$ states of health can be described in this way, and each has been assigned a uility score on the basis of an analysis (Dolan, 1997) of preferences over length and quality of life using data from the general public. Full health is assigned a utility score of 1 and 0 is equivalent to death, with negative values indicating health states worse than death.

Concerns about (lack of) sensitivity and floor/ceiling effects in the standard version recently led to the development of a new version, the EQ-5D-5L. The descriptive system covers the same original five dimensions but the number of levels within each dimension has been extended from three to five (no problems, slight problems, moderate problems, severe problems, extreme problems). In addition, some of the wording has been modified to aid consistency and understanding.[1] The maximum number of health states that can be de-

[1]See the EuroQol website `http://www.euroqol.org/eq-5d-products/how-to-obtain-eq-5d.html` for examples of the question wording used in EQ-5D-3L and EQ-5D-5L.

scribed with the new version is $5^5 = 3125$. Several studies have reported better measurement properties in moving from the EQ-5D-3L to EQ-5D-5L in both specific patient and general population samples (Pickard et al., 2007; Janssen et al., 2008a,b, 2013; Scalone et al., 2013; Agborsangaya et al., 2014; Jia et al., 2014; Pattanaphesaj and Thavorncharoensap, 2015). Utility value sets for EQ5D-5L have been released for England (Devlin et al., 2016), Japan (Ikeda et al., 2015), Canada (Xie et al., 2016) and Uruguay (Augustovski et al., 2015) and similar work is underway in many other countries.

Many studies now include EQ-5D-5L instead of the standard version. Since these studies will form part of the evidence in future economic evaluations, it is important to assess the likely consequences for economic evaluation decisions of moving across the two different versions of EQ-5D, and to develop a basis for using the very large stock of existing evidence based on the 3-level version. In this paper we specify a joint model of the responses to EQ-5D-3L and EQ-5D-5L which allows us to map coherently from EQ-5D-3L to EQ-5D-5L and the reverse. We apply the model to investigate the consistency of the responses to the two descriptive systems, the implied differences in the utility values and the impact on economic evaluation decisions of moving between the two versions in a representative decision problem.

We begin in section 2 by describing the North American dataset we use for the EQ-5D-3L and EQ-5D-5L comparison – one of the few datasets available in which both variants of the instrument are observed in the same survey instrument. In sections 3 and 4, we develop a new flexible modelling approach for analysing the 3-level and 5-level data and report the results of its application. Section 5 uses the results to carry out mapping from 3-level to 5-level and *vice versa*, and section 6 shows the impact of the switch from 3-level to 5-level EQ-5D on the outcome of a representative evaluation of four competing drug therapies for rheumatoid arthritis. Our evidence suggests that a decision by policy-makers to move from EQ-5D-3L to EQ-5D-5L will raise significant doubts about the reliability of many past decisions.

2 The NDBRD dataset

The National Data Bank for Rheumatic Diseases (NDBRD) is a register of patients with rheumatoid disease, primarily recruited by referral from US and Canadian rheumatologists. Information supplied by participants is validated by direct reference to records held by hospitals and physicians.[2] Full details of the recruitment process are given by Wolfe and Michaud (2011). The EQ-5D responses and other patient-supplied data are collected by various means, primarily postal and web-based questionnaires completed directly by patients. Data collection began in 1998 and continues to the present, in waves administered in January and July of each year. In 2011, there was a switch from 3-level to the 5-level version of EQ-5D and both versions were collected in parallel during the January 2011 wave, to allow the effects of the switch to be accommodated in analyses spanning the whole period. Our principal aim is to use data from that wave of the survey to estimate a joint model of the 3- and 5-level responses, which can then be used to map from 3- to 5-level EQ-5D during the pre-2011 period and from 5- to 3-level EQ-5D after January 2011. It then becomes possible to investigate the consistency of the two versions of EQ-5D and assess the impact of mapping between them.

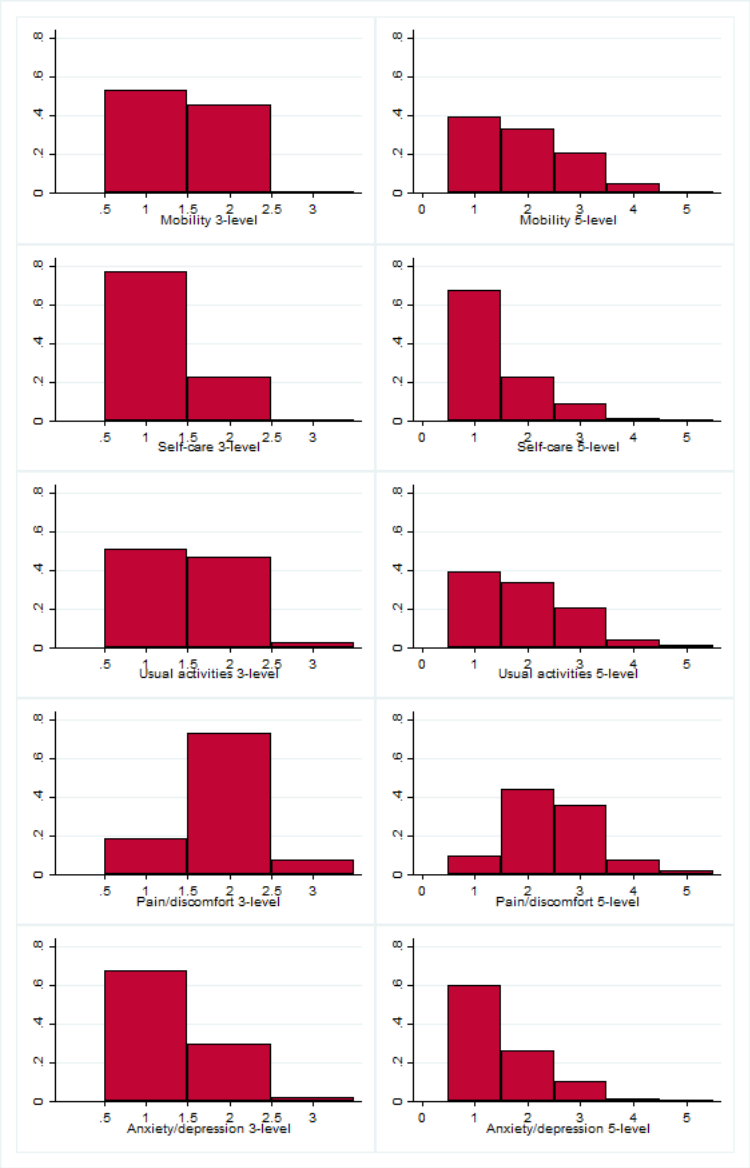
2.1 EQ-5D response distributions

Figure 1 shows histograms of the NDBRD sample response distributions for the 3- and 5-level versions of each domain of EQ-5D. There are clear differences between the distributional shapes for different domains: self-care and anxiety/depression have a dominant mode at the first category; the mobility and usual activities domains also have a decreasing profile but with a heavier central section, while the pain/discomfort domain shows a strong mode in the centre of the distribution. This variation in the shape of the component distributions

[2] A minority of cases come by self-referral, with medical details obtained by NDB in the same way.

underlines the need to use a suitably flexible model specification to analyse the relationship between variants of EQ-5D.

Figure 1: Response histograms for EQ-5D-3L and EQ-5D-5L
(Jan 2011 wave of NDBRD, $n = 5192$)



2.2 Utility scores

For each possible combination of EQ-5D responses, there is a utility value which allows overall health-related quality of life to be estimated and compared across individuals and

conditions. We use the value sets produced by Dolan (1997) and Devlin et al. (2016) for the 3- and 5-level versions of the instrument which, at present, are the standard choices for QALY measurement in England. Dolan (1997) used data from a representative sample of the UK population (2,977 respondents). Each respondent valued 13 hypothetical health states using the time trade-off (TTO) method, generating valuations for a subsample of 42 of the 243 health states described by the EQ-5D-3L. The data were then modelled using regression methods to impute utility values for the remaining health states. Devlin et al. (2016) used a sample of the English population (996 respondents) who valued 10 health states using a composite TTO approach, and 7 paired comparisons of health states via discrete choice experiment tasks. The model selected for the EQ-5D-5L value set for England was a hybrid model using both sets of data (Feng et al., 2016).

Figure 2 shows kernel density estimates of the distributions of utility scores in the NDBRD data, aggregated across all five domains. The distribution is smoother for the 5-level version, particularly towards the top of the range, and this finer structure is a major reason for its adoption in practice. The distribution of utility scores for the 3-level version of EQ-5D has two particularly worrying features. There are ranges with probability mass at or close to zero, particularly around 0.8-1.0 and 0.3-0.45. Consequently, methods for mapping to and from EQ-5D-3L which implicitly assume a smooth positive density can give very poor results (Hernández-Alava et al., 2012). The second striking feature of the distribution for EQ-5D-3L is the large group of cases with utility values close to zero, implying a non-negligible proportion of patients with rheumatoid arthritis (RA) who are in a state comparable to, or worse than, death. The outcomes of evaluation studies often rest on the ability of a therapy to improve the quality of life (QoL) of patients in very poor health states, so the (perhaps implausibly) large frequency of such cases is a potential source of bias in NICE recommendations.

Figure 2: Smoothed empirical distribution functions of EQ-5D-3L and EQ-5D-5L (Jan 2011 wave of NDBRD, $n = 5192$)

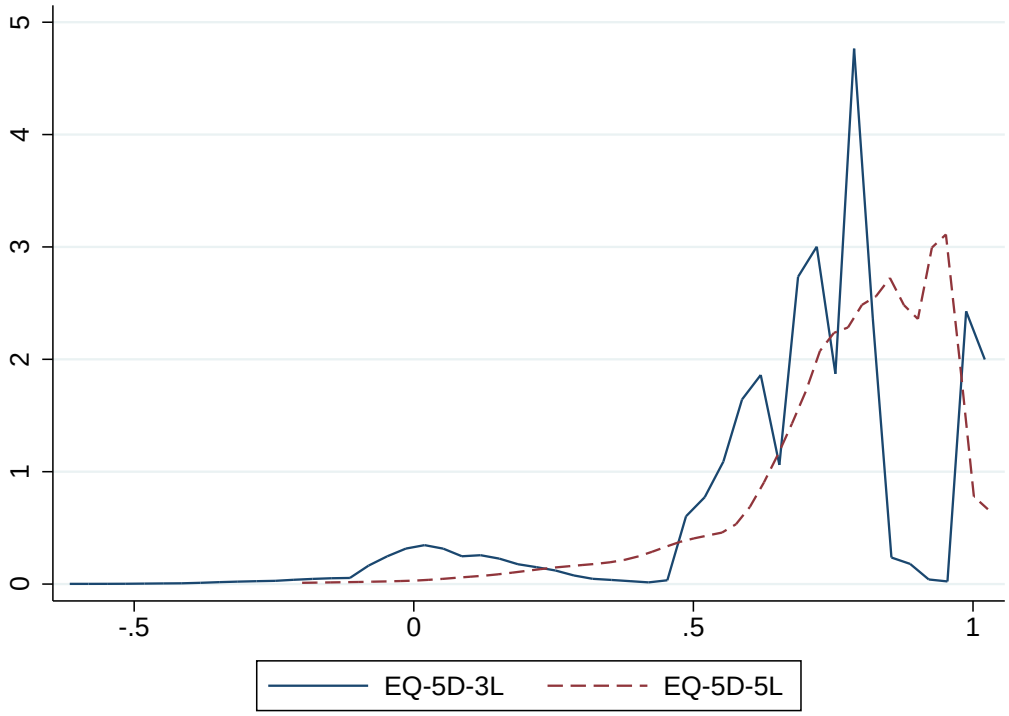


Table 1 summarises the January 2011 NDBRD data on the value scores for the two variants of EQ-5D in terms of their correlation with each other, with basic demographic characteristics, and with a set of clinical outcome measures. They show a high correlation between the two variants of EQ-5D, but the 5-level version has greater sensitivity, since correlations with demographics and clinical outcomes (in the lower panels of Table 1) are uniformly higher for EQ-5D-5L.

Table 2 shows that there is a systematic difference in the 3-level and 5-level utility scores, with the old system generating utilities averaging (in the NDBRD data) only 88% of the utility values given by the new system. This alone could make a significant difference to some evaluation results. It would be inadvisible to address the issue with a simple proportional adjustment, since the ratio of mean scores is not constant but decreases as both general

Table 1: Spearman correlations of 3- and 5-level EQ-5D
(Jan 2011 wave of NDBRD, $n = 4856$)

Variable	EQ-5D-3L	EQ-5D-5L
EQ-5D-3L	1.000	0.849
EQ-5D-5L	0.849	1.000
Female	-0.054	-0.072
Age	0.030	0.055
HAQ score (0-3)	-0.735	-0.766
Pain scale (0-10)	-0.707	-0.711
Overall RADAI score	-0.737	-0.753
Global severity (0-10)	-0.698	-0.726
Disease duration (months)	-0.053	-0.067
Polysymptomatic distress scale	0.462	0.487
Fatigue scale (0-10)	-0.633	-0.670
Sleep disturbance scale (0-10)	-0.506	-0.540
Arthritis activity (general)	-0.611	-0.630
Arthritis activity (today)	-0.672	-0.678
RADAI joints (score)	-0.641	-0.653
RADAI joints (count)	-0.581	-0.593
Morning stiffness (0-6)	-0.538	-0.559
Co-morbidity index (0-9)	-0.344	-0.362
Physical component score (SF-6D)	0.727	0.767
Mental component score (SF-6D)	0.475	0.523
Health satisfaction (0-4)	-0.638	-0.671

severity and pain increase, so the differences are minor at the top end of EQ-5D and much larger at the bottom. Table 2 gives means classified by levels of general disability (in three groups, scores 0-1, 1-2 and 2-3) and pain (in five groups 0-2, 2-4, 4-6, 6-8 and 8-10), as classified by the Stanford Health Assessment Questionnaire (HAQ). The HAQ is widely used by clinicians to measure treatment outcomes; see Bruce and Fries (2003) for a review.

3 A copula model with mixture marginals

Define $Y_{3id} \in \{0, 1, 2\}$ and $Y_{5id} \in \{0, 1, 2, 3, 4\}$ to be the reported outcomes for the dth domain ($d = 1 \dots 5$) of the 3- and 5-level forms of EQ-5D. The model is a system of ten latent regressions, arranged in the five domain groups, with domain d containing the equations for

Table 2: Means of EQ-5D-3L and EQ-5D-5L utility scores by severity of condition (Jan 2011 wave of NDBRD, $n = 5192$)

		3L	5L	Ratio
Overall		0.68	0.77	0.88
By general severity (HAQ) and pain scale category				
General[1]	Pain[2]	3L	5L	Ratio
1	1	0.87	0.92	0.95
1	2	0.76	0.86	0.89
1	3	0.72	0.82	0.88
1	4	0.67	0.78	0.87
1	5	0.51	0.72	0.71
2	1	0.74	0.81	0.91
2	2	0.66	0.75	0.88
2	3	0.60	0.72	0.83
2	4	0.52	0.64	0.81
2	5	0.30	0.53	0.56
3	1	0.63	0.70	0.90
3	2	0.54	0.64	0.84
3	3	0.45	0.56	0.81
3	4	0.35	0.48	0.73
3	5	0.15	0.36	0.42

[1] Groups corresponding to HAQ scores (1) [0-1); (2) [1-2) and (3) [2-3]
[2] Groups corresponding to pain scores (1) [0-2); (2) [2-4); (3) [4-6); (4) [6-8) and (5) [8-10]

Y_{3id} and Y_{5id}:

$$\left.\begin{array}{rcl} Y^*_{3id} & = & X_{id}\beta_{3d} + U_{3id} \\ Y^*_{5id} & = & X_{id}\beta_{5d} + U_{5id} \end{array}\right\} \qquad d = 1...5 \tag{1}$$

where i indexes individual cases and we assume random sampling so that all sampled variables are independent across individuals. $X_{i1}...X_{i5}$ is a collection of row vectors of covariates and β_{3d}, β_{5d} are column vectors of cefficients conformable with X_{id}. We assume that the covariate vector X_{id} is the same for both the 3-level and 5-level version of the rth domain, but may differ between domains. U_{3id}, U_{5id} are unobserved residuals which may be stochastically dependent and non-normal. The latent dependent variables Y^*_{3id}, Y^*_{5id} are not observed directly but they have observable ordinal counterparts, Y_{3id}, Y_{5id}, generated by the following threshold-crossing conditions:

$$Y_{kid} = q \quad \text{iff} \quad \Gamma_{kqd} \leq Y^*_{kid} < \Gamma_{k(q+1)d}\ ; \qquad q = 1...Q_k\ ;\ k = 3, 5 \tag{2}$$

where $Q_k = 3$ or 5 is the number of categories of Y_{kid} and the Γ_{kqd} are threshold parameters, with $\Gamma_{k1d} = -\infty$ and $\Gamma_{k(Q_k+1)d} = +\infty$.

We decompose the residual U_{kid} into a single between-group factor which represents the individual's general tendency to give more or less positive responses and a specific residual correlated within but not between domains:

$$U_{kid} = \psi_{kd} V_i + \varepsilon_{kid} \tag{3}$$

where the ψ_{kd} are a set of ten parameters. Supressing the i subscript, we make the following standard assumptions:

$$V \perp\!\!\!\perp \varepsilon_{kd} | \boldsymbol{X} , \qquad k = 3, 5, \;\; d = 1 \ldots 5 \tag{4}$$

$$\varepsilon_{kd} \perp\!\!\!\perp \varepsilon_{lg} | \boldsymbol{X} , \qquad k = 3, 5, \;\; d \neq g \tag{5}$$

$$\varepsilon_{3d} \not\perp\!\!\!\perp \varepsilon_{5d} | \boldsymbol{X} , \qquad d = 1 \ldots 5 \tag{6}$$

where $\boldsymbol{X} = [X_1 \ldots X_5]$ and $\perp\!\!\!\perp$ and $\not\perp\!\!\!\perp$ denote statistical independence and (possible) dependence respectively.

3.1 Within-domain variation

We use a 1-parameter copula representation to capture the dependence between the 3-level and 5-level responses for any domain. When applying this approach in a single domain, we assume that $V_i = 0$ almost surely.[3]. Define $F_d(\varepsilon_{3d}, \varepsilon_{5d})$ to be the distribution function (df) for domain d and $F_{3d}(\varepsilon_{3d}) = F_d(\varepsilon_{3d}, \infty)$ and $F_{5d}(\varepsilon_{5d}) = F_d(\infty, \varepsilon_{5d})$ to be the marginals. The joint residual df for domain d is specified as:

$$F_d(\varepsilon_{3d}, \varepsilon_{5d}) = c_d \left(G_{3d}(\varepsilon_{3d}), G_{5d}(\varepsilon_{5d}); \theta_d \right) \tag{7}$$

[3]This is essentially equivalent to using the copula representation for the whole residuals U_{3d}, U_{5d} rather than $\varepsilon_{3d}, \varepsilon_{5d}$

where $G_{kd}(.)$ is the marginal df of ε_{kd} and θ_d is a parameter controlling the pattern of dependence between ε_{3d} and ε_{5d}. The copula function $c_d : [0,1] \times [0,1] \to [0,1]$ has the properties $c_d(0,u) = c_d(u,0) = 0$ and $c_d(1,u) = c_d(u,1) = u$ for any $u \in [0,1]$ (Trivedi and Zimmer, 2005). We consider the following candidate forms:

Gaussian: $c(\varepsilon_3, \varepsilon_5) = \Phi\left(\Phi^{-1}(\varepsilon_3), \Phi^{-1}(\varepsilon_5); \theta\right)$

where $\Phi(.,.;\theta)$ is the distribution function of the bivariate normal with correlation coefficient $-1 \leq \theta \leq 1$ and $\Phi^{-1}(.)$ is the inverse of the univariate $N(0,1)$ df

Clayton:

$$c(\varepsilon_3, \varepsilon_5) = \begin{cases} \left[\max\left\{\varepsilon_3^{-\theta} + \varepsilon_5^{-\theta} - 1, 0\right\}\right]^{-1/\theta} & \text{for } 0 < \theta \leq \infty \\ \varepsilon_3 \varepsilon_5 & \text{for } \theta = 0 \end{cases}$$

Frank:

$$c(\varepsilon_3, \varepsilon_5) = \begin{cases} -\frac{1}{\theta} \ln\left(1 + \frac{\left(e^{-\theta\varepsilon_3} - 1\right)\left(e^{-\theta\varepsilon_5} - 1\right)}{e^{-\theta} - 1}\right) & \text{for } \theta \neq 0 \\ \varepsilon_3\, \varepsilon_5 & \text{for } \theta = 0 \end{cases}$$

Gumbel: $c(\varepsilon_3, \varepsilon_5) = exp\left(-\left[(-\ln \varepsilon_3)^\theta + (-\ln \varepsilon_5)^\theta\right]^{1/\theta}\right)$ for $\theta \geq 1$

Joe: $c(\varepsilon_3, \varepsilon_5) = 1 - \left[(1-\varepsilon_3)^\theta + (1-\varepsilon_5)^\theta - (1-\varepsilon_3)^\theta (1-\varepsilon_5)^\theta\right]^{1/\theta}$ for $\theta \geq 1$

These are capable of representing a range of dependence structures. The Gaussian and Frank copulas are similar in the sense that both allow for positive and negative dependence and dependence is symmetric in both tails. However, compared to the Gaussian copula, the Frank copula generates dependence weaker in the tails and stronger in the centre of the distribution. The Clayton, Gumbel and Joe copulas allow only positive dependence, and dependence in the tails is asymmetric. The Clayton copula exhibits strong left tail dependence and relatively weak right tail dependence. Thus, if two variables are strongly correlated at low values but not so correlated at high values, then the Clayton copula is a good choice. The Gumbel and Joe copulas display the opposite pattern with weak left tail

dependence and strong right tail dependence. Right tail dependence is stronger for the Joe than the Gumbel copula.

The within-domain specification is completed by a normal mixture assumption which allows the residuals to have a non-normal form:

$$G(\varepsilon) = \pi\Phi\left((\varepsilon - \mu_1)/\sigma_1\right) + [1-\pi]\Phi\left((\varepsilon - \mu_2)/\sigma_2\right) \tag{8}$$

where: $0 \leq \pi \leq 1$ is the mixing parameter; (μ_1, μ_2) and $(\sigma_1, \sigma_2 \geq 0)$ are location and dispersion parameters constrained to satisfy the mean and variance normalizations $\pi\mu_1 + (1-\pi)\mu_2 \equiv 0$ and $\pi\left(\sigma_1^2 + \mu_1^2\right) + (1-\pi)\left(\sigma_2^2 + \mu_2^2\right) = 1$. These normal mixtures are able to capture a wide range of distributional shapes, especially skewness and bimodality. The mixture form (8) can be implemented with various degrees of generality, by assuming the same parameter values $(\pi, \mu_1, \mu_2, \sigma_1, \sigma_2)$ for all residuals, or allowing them to vary with domain $d = 1...5$ and/or EQ-5D level $k = 3, 5$.

Conditional on X_d, the probability of observing any values $Y_{3d} = y_q$ and $Y_{5d} = r$ is:

$$\begin{aligned} P\left(q, r|X_d\right) &= c_d\left(G_{kd}(q+1), G_{kd}(r+1)\right) - c_d\left(G_{kd}(q+1), G_{kd}(r)\right) \\ &\quad -c_d\left(G_{kd}(q), G_{kd}(r+1)\right) + c_d\left(G_{kd}(q), G_{kd}(r)\right) \end{aligned} \tag{9}$$

where G_{kd} is shorthand for $G_{kd}\left(\Gamma_{kqd} - X_d\beta_{kd}\right)$. Thus, calculating the likelihood for a domain-specific bivariate model requires four copula evaluations for each sample observation.

3.2 Between-domain variation

In high-dimensional ordinal-variable applications, copula models can become infeasible. In our application with 10 paired domain indicators, the joint likelihood is the probability of a 10-dimensional hyper-rectangle $[a_{0,1}, a_{1,1}) \times \ldots \times [a_{0,10}, a_{1,10})$. Conditional on $X = X_1 \ldots X_5, V$,

this probability can in principle be constructed from a 10-dimensional copula $C(.)$:

$$P\left(Y_{31}, Y_{51}, \ldots, Y_{35}, Y_{55}|X, V\right) = \sum_{j_1=0}^{1} \ldots \sum_{j_{10}=0}^{1} (-1)^{j_1+\ldots+j_{10}} C\left(G_{31}(a_{j_1,1}), \ldots, G_{55}(a_{j_{10},10})\right) \quad (10)$$

This could require $2^{10} = 1024$ evaluations of the copula, which is both time-consuming and vulnerable to build-up of truncation error in finite arithmetic. Possible solutions to the problem work by imposing structure on the copula, building it up from bivariate sub-copulas. The most convenient of these methods are based on vine structures Bedford and Cooke (2002), particularly the specific D-vine form (Panagiotelis et al., 2012).

However, the D-vine structure is most convincing when there is a natural ordering of the observed variables, particularly temporal sequencing (as in the application by Panagiotelis et al. (2012) to a sequence of four observations on headache spaced through the day). In our case, although the component items of EQ-5D-5L were asked in sequence and then the items of EQ-5D-3L later in the questionnaire, this ordering of items does not correspond at all to the natural connections between the 3-level and 5-level items through their shared inherent meaning. For that reason, we adopt a different approach, using five separate bivariate copulas for the five domains of EQ-5D, and connecting those domains via the latent factor V which represents the respondent's background response behaviour. In the most general specification, we allow V to have a two-part normal mixture distribution with density $\frac{p}{s_1}\phi\left([V - m_1]/s_1\right) + \frac{1-p}{s_2}\phi\left([V - m_2]/s_2\right)$ where $0 \leq p \leq 1$, $s_1, s_2 \geq 0$ and m_1, m_2 are parameters.

The joint distribution of $(Y_{31}, Y_{51}) \ldots (Y_{35}, Y_{55})$ is:

$$Pr\left(Y_{31}, Y_{51} \ldots Y_{35}, Y_{55}|\boldsymbol{X}\right) = \int \prod_{d=1}^{5} P(Y_{3d}, Y_{5d}|X_d, v) \left[\frac{p}{s_1}\phi\left(\frac{v - m_1}{s_1}\right) + \frac{1-p}{s_2}\phi\left(\frac{v - m_2}{s_2}\right)\right] dv \quad (11)$$

We use Gauss-Hermite quadrature to evaluate the integral in (11) at each observation to give the likelihood function.

4 Modelling results

Our aim is to estimate the joint distribution of the responses to the 3L and 5L variants of the EQ-5D survey instrument, conditional on demographic characteristics (age and gender), and clinical measures of the severity of the underlying rheumatic condition. We use seven covariates: age, gender, the HAQ disability score, the pain scale, and the squares and product of the HAQ and pain scales.

The HAQ is based on patient self-reporting of the degree of difficulty experienced over the previous week in eight categories: dressing and grooming, arising, eating, walking, hygiene, reach, grip, and common daily activities. It is widely used by clinicians to measure health outcomes. It is scored in increments of 0.125 between 0 and 3 (although it is standard to consider it fully continuous), with higher scores representing greater degrees of functional disability. The HAQ instrument also includes separately a patient self-report of pain scored on a Visual Analogue Scale (0-10).

4.1 Domain-specific modelling

We first examine each of the five domains of EQ-5D separately using a bivariate approach, implemented in the Hernández-Alava and Pudney (2016) Stata `bicop` routine. Table 3 summarises the sample fit of alternative copula functions for the 3- and 5-level variants for each of the five domains, where we retain the standard assumption of Gaussian marginals. There is no single best choice of copula: the Gaussian form fits best for dimensions 1 and 3 (mobility and usual activities), the Frank copula fits best for dimensions 2 and 5 (self-care and anxiety/depression) while the Gumbel copula fits best for the pain/discomfort dimension. This coincides with differences in the empirical distributions of Figure 1 between these three groups of domains. The Frank copula (which allows weaker dependence in the tails than

the centre of the distribution) works better than the Gaussian copula when the tails of the response distribution are relatively heavy. The Gumbel copula which has asymmetric dependence in the tails (stronger dependence at higher values) fits better when there is a central mode and implies different patterns of dependence in both tails of the distribution.

Table 2 also gives the results of the Wald test of the null hypothesis that the coefficient vectors relating the (latent) response to age, gender and disease severity are identical in the 3- and 5-level variants. The hypothesis is clearly rejected for the domains of mobility and pain. This finding shows that the effect of the move to 5 levels is not simply a uniform re-alignment of the response level.

Table 3: Sample fit of domain-specific models for alternative copula functions with Gaussian marginals)

	Copula				
	Gaussian	Frank	Clayton	Gumbel	Joe
Mobility domain					
Log-likelihood	**-6656.54**	-6665.73	-6727.46	-6669.82	-6736.73
$\chi^2(7)$ for $H_0 : \beta_3 = \beta_5$	29.02***	29.49***	23.82***	33.64***	37.14***
Self-care domain					
Log-likelihood	-4221.35	**-4212.35**	-4248.89	§	§
$\chi^2(7)$ for $H_0 : \beta_3 = \beta_5$	8.31	5.98	5.35		
Usual activities domain					
Log-likelihood	**-6772.96**	-6796.04	-6866.11	-6785.64	-6829.65
$\chi^2(7)$ for $H_0 : \beta_3 = \beta_5$	10.87	10.22	10.89	11.23	11.53
Pain/discomfort domain					
Log-likelihood	-6148.63	-6148.07	-6190.84	**-6147.80**	-6199.63
$\chi^2(7)$ for $H_0 : \beta_3 = \beta_5$	29.75***	30.26***	32.71***	29.09***	26.82***
Anxiety/depression domain					
Log-likelihood	-6243.59	**-6238.86**	-6300.55	-6244.72	-6302.70
$\chi^2(7)$ for $H_0 : \beta_3 = \beta_5$	12.05*	8.56	5.10	10.66	11.86

Best-fitting models in bold type (all models have 15 parameters). Statistical significance: * = 10%, ** = 5%, *** = 1%. § No convergence.

We also explored the possibility of non-normality using a 2-part Gaussian mixture for each residual. The assumption of normal marginals was acceptable in terms of the Akaike (AIC) and Bayesian (BIC) information criteria for the mobility, self-care and anxiety/depression

domains, but there was significant evidence of modest departures from normality for the usual activities and pain/discomfort domains. Table 4 summarises the preferred specifications for those two domains, comparing them with the simpler Gaussian-marginal models. Note that the conclusions about the equality of coefficients for these two dimensions are not affected by the non-normality of the residual distributions. Figure 3 plots the residual distributions for these two dimensions and compares them to the N(0,1) distribution. The residual distributions for the usual activities dimension and for the EQ-5D-5L pain/anxiety dimension are similar, both with a fatter right tail of the distribution. The residual distribution for the EQ-5D-5L pain/anxiety dimension departs from normality with a much bigger central mode consistent with its unique distributional shape in Figure 1.

Table 4: Non-normality in residual distributions

	Gaussian marginals		Non-Gaussian marginals			
Domain	AIC	BIC	Preferred mixture specification	AIC	BIC	Coefficient equality test: $\chi^2(7)$
Usual activities[1]	13587.9	13725.5	equal	13550.5	13707.8	8.39
Pain/discomfort[2]	12337.6	12475.3	unequal	12252.9	12429.9	40.91***

Statistical significance: * = 10%, ** = 5%, *** = 1%.. [1] Gaussian copula. [2] Gumbel copula.

4.2 Joint modelling of all domains

We now examine the joint model. Table 5 summarises the sample fit of alternative joint models[4]. All of them are based on the best fitting copulas for each dimension found in the last section (the Gaussian copula for mobility and usual activities, Frank for self-care and anxiety/depression and Gumbel for pain/discomfort). Model (a) is the baseline model with no mixtures in ε; model (b) allows a common mixture, constrained to be the same for the residuals in all 10 equations; and model (c) allows for one common mixture for the usual

[4]The likelihood functions are calculated using Gauss-Hermite quadrature with 15 integration points. Differences in the parameter estimates are negligible when varying the number of integration points.

Figure 3: Residual distribution for the usual activities domain
(constrained equal for 3- and 5-level residuals)

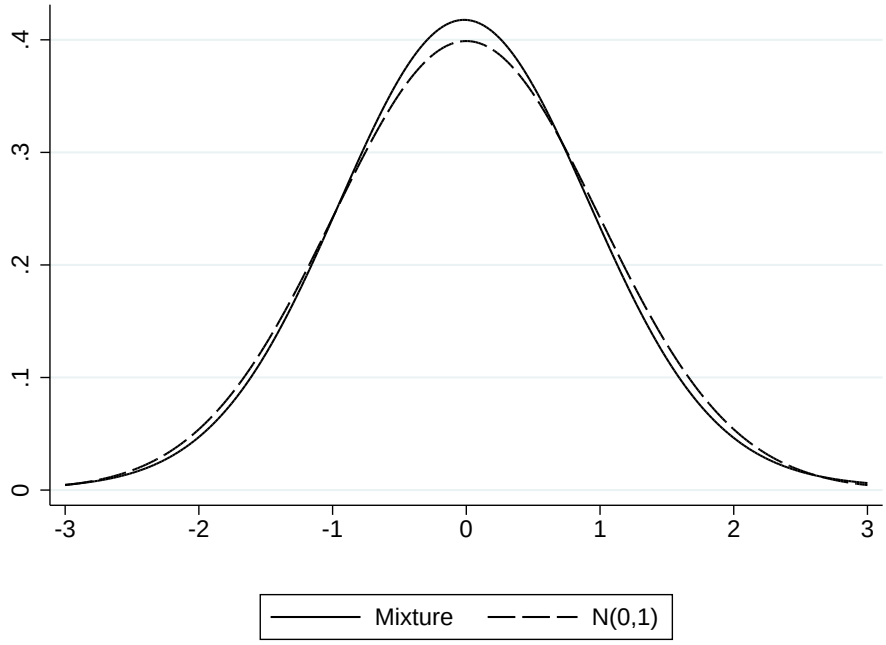
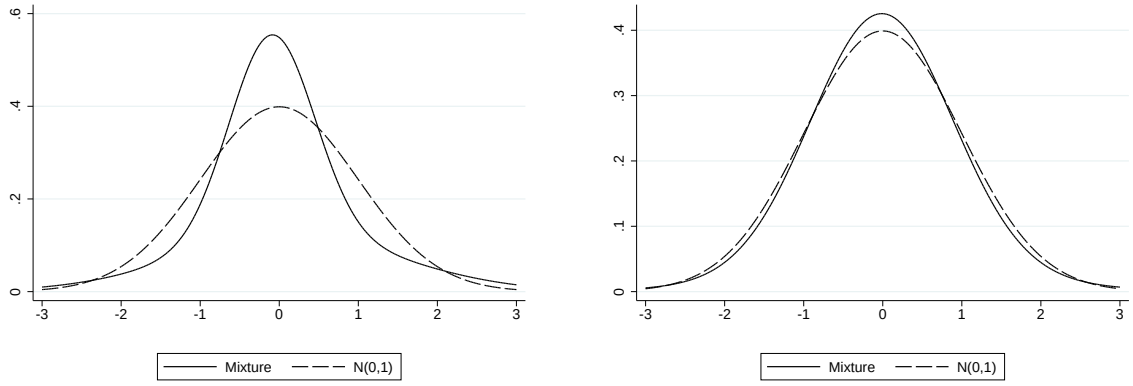


Figure 4: Unconstrained residual distributions for the pain/discomfort domain



(a) 3-level **(b)** 5-level

activities domain and 2 unequal mixtures for the pain/discomfort domain, following the pattern of the domain-specific results. The joint log-likelihood, AIC and BIC for the model with independent EQ-5D dimensions are -29958.431, 60144.86 and 60892.12 respectively, indicating that the joint model provides a better fit to the data. The joint model with a common mixture, model (b), gives the best fit to the data according to AIC and BIC.

The conclusions about the equality of coefficients are not affected by the choice of residual distributions and are in line with the conclusions of the domain-specific bivariate models. The estimated coefficients of the domain-specific bivariate and joint models can be found in Appendix Table A1.

Table 5: Sample fit of joint copula models

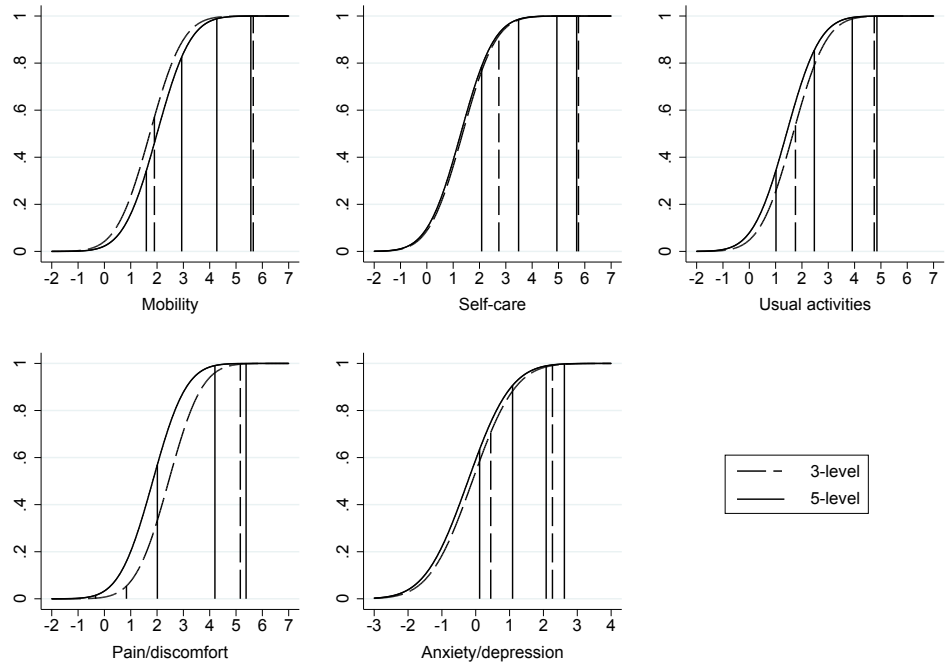
	Type of mixture in ε		
	(a) None	(b) Equal	(c) Unequal
Log-likelihood	-29197.46	-29136.23	-29132.50
Number of parameters	115	118	124
AIC	58624.91	58508.46	58513.00
BIC	59378.73	59281.93	59325.80
Coefficient equality			
Mobility domain			
Equality of β $\chi^2(7)$	26.59***	26.53***	25.69***
Equality of ψ $\chi^2(1)$	0.18	0.29	0.00
Equality of β and ψ $\chi^2(8)$	28.59***	26.53***	28.73***
Self-care domain			
Equality of β $\chi^2(7)$	4.14	3.50	3.99
Equality of ψ $\chi^2(1)$	3.02*	3.37*	4.17**
Equality of β and ψ $\chi^2(8)$	9.60	8.91	10.80
Usual activities domain			
Equality of β $\chi^2(7)$	8.81	7.93	9.39
Equality of ψ $\chi^2(1)$	0.33	0.21	0.45
Equality of β and ψ $\chi^2(8)$	12.77	10.82	11.88
Pain/discomfort domain			
Equality of β $\chi^2(7)$	31.64***	30.19***	36.58***
Equality of ψ $\chi^2(1)$	18.80***	21.42***	29.27***
Equality of β and ψ $\chi^2(8)$	46.98***	50.65***	66.01***
Anxiety/depression domain			
Equality of β $\chi^2(7)$	9.27	8.70	9.36
Equality of ψ $\chi^2(1)$	2.68	2.75*	3.75*
Equality of β and ψ $\chi^2(8)$	11.07	10.54	11.99

Statistical significance: * = 10%, ** = 5%, *** = 1%.

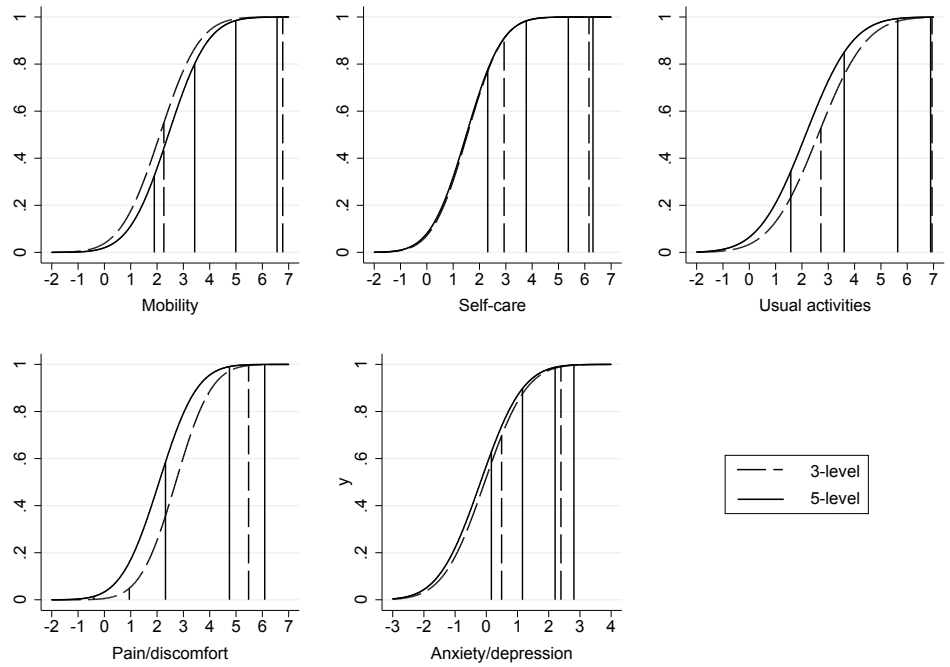
Figure 5 illustrates the effect of the differences in the distribution functions (df) of the latent variables evaluated at the average across sample values of the covariates. For both models the df's of the underlying latent variables of the 3- and 5-level EQ-5D in the self-care and anxiety/depression domains are almost identical. Moreover, the position of the

two thresholds of the 3-level version are consistent with the idea of a re-alignment of the response levels, where the first and the second thresholds of the 3-level version fall respectively between the first and second, and third and fourth thresholds in the 5-level version. The df's in the usual activities domain are very similar in the domain-specific model; athough slightly less so in the joint model, the differences are not statistically significant. For the mobility and pain/discomfort domains, the differences portrayed in the graphs are sizeable and statistically significant in both models. The pain/discomfort domain displays the most noticeable difference between the df's. The mobility domain is unique in that the df of the 3-level version lies to the left of the df of the 5-level version, the reverse is true for all other domains.

Figure 5: Domain comparisons: distribution functions of the latent variables evaluated at the average across sample values of the covariates



(a) Domain-specific models



(b) Joint model

5 Mapping

The best method of mapping between alternative preference-based measures depends on the nature of the cost-effectiveness study in which the measure is to be used. Suppose, for example, that the study is to be done on the new 5-level basis, but the available evidence comes from a clinical trial in which the older EQ-5D-3L scale is measured. The key concept is the mean QALY, which should be constructed as $E\{Q(\upsilon_5(Y_5))\}$, where $E\{.\}$ is the expectation with respect to whatever population is potentially affected by the treatment.

There are two technical issues to be considered in mapping from 3L evidence to 5L-based evaluation. First, the form of the function, $Q(.)$, which maps utilities into QALYs. In most evaluation studies, the QALY calculation $Q(.)$ is a linear function of the utilities, so that $E\{Q(\upsilon_5(Y_5))\} = Q(E\{\upsilon_5(Y_5)\})$. In other words, we can simply predict the utility outcome $\upsilon(Y_5)$ and use that prediction in calculating QALYs. If the predictor is an unbiased (or consistent) estimator of $E[\upsilon(Y_5)]$, it will give an unbiased (consistent) evaluation of the expected QALY.

The second issue is the choice of predictor for $\upsilon(Y_5)$. We have argued here that a predictor based on a full model of $Pr(Y_5|Y_3, X)$ uses more information and is capable of giving better results than the alternative approach to mapping, which attempts to model $E(\upsilon_5(Y_5)|\upsilon_3(Y_3), X)$ directly – often using methods like linear regression which are not well suited to the non-standard distributions involved. When using our approach, it is important to realise that the utility scales $\upsilon(.)$ are nonlinear functions of the vector Y, so $E(\upsilon_5(Y_5)) \neq \upsilon_5(E[Y_5])$. We should not map the observed 3-level health description Y_3 into the 5-level descriptive system Y_5 and then apply the utility scale $\upsilon_5(.)$. Instead, the appropriate method is to use the model estimated from NDBRD data to evaluate the probability of each possible configuration of Y_5 conditional on Y_3, X and use those probabilities as weights

to evaluate the conditional expectation of υ. Specifically, the conditional df of the valuation υ_5 is:

$$Pr\left(\upsilon_5(Y_5) \leq \Upsilon | Y_3, X\right) = \sum_{Y_5 \epsilon U_\Upsilon} Pr(Y_5 | Y_3, X) \tag{12}$$

where U_Υ is the set $\{Y_5 : \upsilon_5(Y_5) \leq \Upsilon\}$ and Υ is any given constant. The mean of the distribution is:

$$E\left(\upsilon_5(Y_5) | Y_3, X\right) = \sum_{Y_5 \epsilon S_5} \upsilon_5(Y_5) Pr(Y_5 | Y_3, X) \tag{13}$$

where S_5 is the set of (3125) possible values that the vector Y_5 might take.

In the published literature, several authors have commented on the loss of variation induced by mapping (Rivero-Arias et al., 2010; Brazier et al., 2010; Longworth and Rowen, 2011; Fayers and Hays, 2014). The sample variance of the mean predictor (13) will always be lower than the variance of the unknown true $\upsilon_5(Y_5)$, because the modelling process can only predict variation in $\upsilon_5(Y_5)$ arising from Y_3 and X, not the other "unexplained" components of variation. In standard cases where the QALY calculation is linear in utilities, this does not matter, since only the conditional mean of $\upsilon_5(Y_5)$ is required. If the aim were to estimate the variance of $\upsilon_5(Y_5)$, one would not do it by using the variance of the predictor (13); instead, the appropriate method would be to calculate directly the variance of the estimated distribution (12), which would give a consistent estimate of $var(\upsilon_5(Y_5))$ if the mapping model is correctly specified and estimated.

Both the distribution (12) and its mean (13) can be evaluated at the sample values Y_{i3}, X_i, averaged over the whole sample or a subsample, and then compared with the corresponding empirical df or mean of the directly observed 3-level scores, $\upsilon_3(Y_{i3})$. This can be done empirically for the pre-January 2011 waves of the NDBRD dataset and in reverse (predicting Y_3 conditional on Y_5) for the post-January 2011 waves. Figure 6a uses the set of domain-specific bivariate models (assuming independence across domains) to compare the predictive df $n^{-1} \sum_{i=1}^{n} Pr\left(\upsilon_5(Y_5) \leq \Upsilon | Y_{i3}, X_i\right)$ and the directly-observed empirical

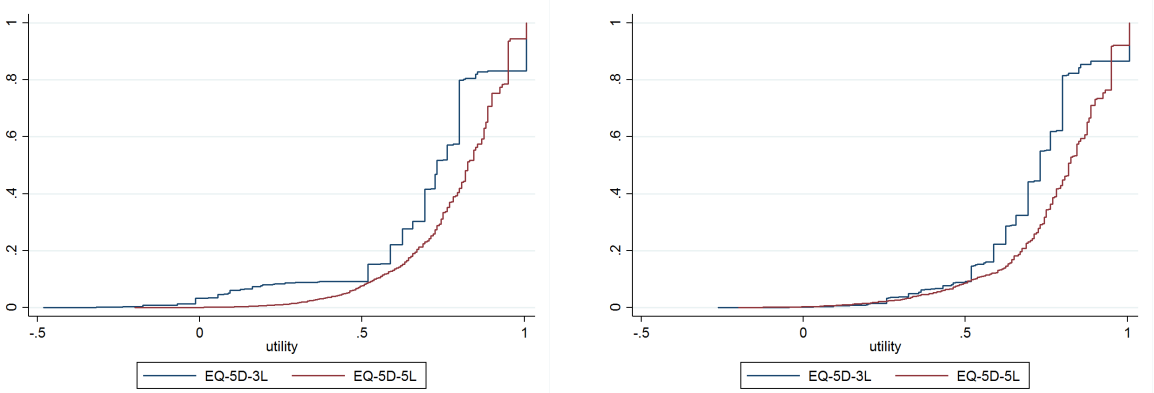
df $n^{-1}\sum_{i=1}^{n}\mathbb{1}\left(\upsilon_3(Y_{i3})\leq\Upsilon\right)$ for the Jan 2010 wave of NDBRD, where $\mathbb{1}(.)$ is the indicator function. Figure 6b makes the reverse comparison of the predictive df for $\upsilon_3(Y_3)$ with the empirical df of $\upsilon_5(Y_5)$ for the Jan 2012 wave. Figure 7 makes the same comparisons for the joint model allowing for between-domain correlation.

There are three striking features of Figures 6 and 7, with important implications for the economic evaluations carried out for public bodies like NICE. First, the predictive and actual distributions of the 5-level variant of EQ-5D are similar and much smoother than the corresponding distributions for the 3-level variant. This is an encouraging finding: if a decision maker elects to recommend the use of the new 5-level instrument and associated scoring, it may be possible to continue to use older 3-level-based evidence with appropriate mapping to 5-level.

Second, the predictive and actual distributions for the 3L variant differ in one important feature: the prediction of EQ-5D-3L from the 5-level responses (Figures 6b or 7b) fails to capture the hump in the directly-observed empirical distribution in the neighbourhood of zero. This is clearly a feature of the 3-level utility value set, rather than a mismatch between the 3-level and 5-level descriptive systems. This finding suggests that it would be inappropriate for decision makers to retain the older 3-level instrument and tariff for economic evaluations, having to rely on mapping data from new trials.

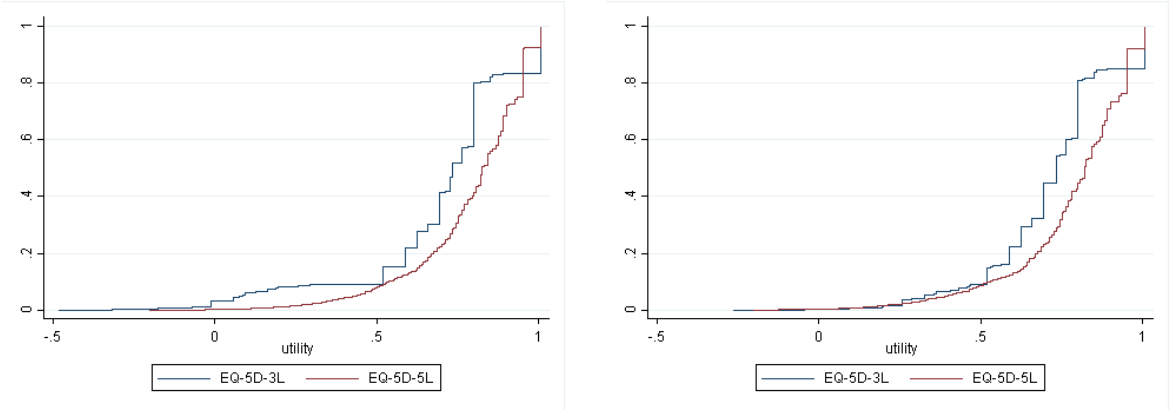
Third, there is a large difference between the 3-level and 5-level distributions of EQ-5D scores, whether directly observed or mapped. Utility scores tend to be systematically higher under the 5-level scoring scheme, so the df for EQ-5D-3L lies entirely to the left of the df for EQ-5D-5L. If no other adjustment were made, this alone might be enough to change many evaluation results, in the absence of offsetting adjustments to the evaluation methodology.

Figure 6: Cross-mapping based on independent domain-specific bivariate models



(a) Jan 2010: 3-level→5-level **(b)** Jan 2012: 5-level→3-level

Figure 7: Cross-mapping based on the joint model with between-domain correlation



(a) Jan 2010: 3-level→5-level **(b)** Jan 2012: 5-level→3-level

Table 6 shows average values of directly-measured $\upsilon_3(Y_3)$ and the prediction $E[\upsilon_5(Y_5)|Y_3, X]$ for the 2010 wave of NDBRD, and of the prediction $E[\upsilon_3(Y_3)|Y_5, X]$ and directly-measured $\upsilon_5(Y_5)$ for the 2012 wave using the joint model. Results are given for the whole sample and subgroups defined in terms of disease severity and demographic characteristics; sample standard deviations of the measured and predicted utilities are are also shown. As expected, there are higher mean values and smaller standard deviations for the EQ-5D-5L scores (whether predicted or directly observed) than for EQ-5D-3L, resulting from the

different scoring of poor health states by the two value sets. Another consequence of this is the much steeper severity gradient for the mean EQ-5D-3L utilities than for EQ-5D.

There is a slight tendency for both the 3-level and 5-level utilities to decline over time as the health states of those individuals who appear in both waves tend to worsen. However, the means of predicted and directly-observed versions of each measure are remakably close both overall and in terms of their severity and demographic profiles.

We also see the anticipated smaller standard deviations of the predicted than directly-observed utilities as a consequence of the use of expected value prediction. This is of no importance for the evaluation described in the next section (since the criterion is based on the mean QALY), but it would be a concern for any evaluation that aims to investigate the distributional pattern of QALY gains within each population group. In that case, appropriate measures constructed from the full distribution (12) would need to be used.

Table 6: Means and standard deviations of actual and predicted (joint model) EQ-5D-3L and EQ-5D-5L by severity of condition, age and gender. (NDBRD. January 2010 wave n = 3877; January 2012 wave n = 3911)

	January 2010		January 2012	
	EQ-5D-3L (actual) mean (SD)	EQ-5D-5L (predicted) mean (SD)	EQ-5D-3L (predicted) mean (SD)	EQ-5D-5L (actual) mean (SD)
Overall	0.70	0.79	0.69	0.78
	(0.25)	(0.16)	(0.21)	(0.19)
Severity group				
Mild	0.88	0.92	0.87	0.92
(HAQ group 1, Pain group 1)	(0.12)	(0.04)	(0.08)	(0.07)
Medium	0.62	0.70	0.61	0.72
(HAQ group 2, Pain group 3)	(0.15)	(0.09)	(0.11)	(0.11)
Severe	0.12	0.39	0.12	0.31
(HAQ group 3, Pain group 5)	(0.30)	(0.15)	(0.19)	(0.21)
Female <65	0.70	0.78	0.68	0.77
	(0.26)	(0.17)	(0.23)	(0.20)
Male <65	0.71	0.80	0.67	0.77
	(0.25)	(0.16)	(0.24)	(0.20)
Female 65-79	0.71	0.79	0.69	0.79
	(0.24)	(0.15)	(0.20)	(0.17)
Male 65-79	0.73	0.82	0.73	0.82
	(0.22)	(0.15)	(0.19)	(0.15)
Female ≥ 80	0.65	0.76	0.66	0.76
	(0.25)	(0.17)	(0.20)	(0.18)
Male ≥ 80	0.74	0.82	0.70	0.79
	(0.17)	(0.12)	(0.17)	(0.16)

6 The impact on cost-effectiveness analysis

We now use a published cost-effectiveness study to examine the potential consequences of moving from EQ-5D-3L to EQ-5D-5L as a basis for economic evaluation. We first replicate the economic evaluation results in Wailoo et al. (2014), which use EQ-5D-3L data collected as part of a trial. Then we repeat the analysis using EQ-5D-5L obtained using the mapping models developed in this paper. Wailoo et al. (2014) estimate the cost-effectiveness of combinations of disease-modifying anti-rheumatic drugs (DMARDs) and short-term adminis-

tration of the steroid prednisolone (PNS), using data from the 2-year CARDERA trial which involved 467 adult patients with early active RA (less than two years of disease duration) in a placebo-controlled factorial design. Two DMARDS were used in the trial, methotrexate (MTX) and ciclosporin (CS). All patients received MTX, half received step-down PNS[5] and half CS, generating four treatment groups: (1) monotherapy (MTX only), (2) combination DMARDs (MTX and CS), (3)DMARD and steroid (MTX and PNS) and (4) triple therapy (MTX, CS and PNS). Further details of the methods and clinical effectiveness can be found in Choy et al. (2008).

The key criterion used in cost-effectiveness analysis is the Incremental Cost-Effectiveness Ratio (ICER), defined as the difference in costs between two different treatment strategies, expressed as a ratio to the difference in the QALYs that they achieve. Treatments with ICERs below a certain threshold are usually considered cost-effective. In the UK, NICE guidance on technology appraisal refers to a specific range £20,000-£30,000 (NICE, 2013), but see also Claxton et al. (2015) who argue for a lower threshold.

Resource use (prescription drugs, hospitalizations, tests, imaging, surgical procedures and community care visits) was directly observed over the two years of the trial and costed using 2011-2012 figures. The mean discounted cost of each treatment strategy is shown in the first row of Table 7. QALY estimates were derived from EQ-5D-3L responses observed at baseline and 6, 12, 18 and 24 months. The discounted QALY total was then estimated as the area under the linear interpolation of the five points. Table 7 presents the mean costs and QALYs after two years for the sample of patients with complete data (n=241).[6] Of all four treatment strategies, triple therapy is the least costly and most effective (higher QALYs), thus dominating all other strategies. Of the remaining three treatment strategies,

[5]Initially dosed at 60mg/day, reducing to 7.5mg/day at 6 weeks and stopped by 34 weeks.

[6]Note that there are minor differences between the numbers reported in Table 7 and those in Wailoo et al. (2014) due to missing data in the variables used to predict EQ-5D-5L for one patient, but results are unaffected.

the DMARD combination is dominated by a DMARD plus steroid, being more costly and less effective. Monotherapy is more effective but also more costly than MTX plus steroid, with an ICER of £13,714 which lies comfortably below a conventional cost-effectiveness threshold of (say) £20,000 per QALY.

We then repeated the estimation of QALYs using EQ-5D-5L predicted from the models developed in this paper, conditional on the EQ-5D-3L responses observed in the trial. This was done separately at each 6-monthly observation of EQ-5D-3L and the total computed by interpolation as before. Note that, since this construction is a linear function of the EQ-5D responses Y, our use of $E(Y_5|Y_3, X)$ as a predictor does not introduce bias into the QALY evaluation, as it would for a nonlinear function of Y. The independent domains model and the more complex joint model give very similar results in terms of total QALYs but the point estimates of the joint model seem to fall below those of the independent domains model. The mapped EQ-5D-5L QALYs are consistently larger (by 15-25%) than the EQ-5D-3L QALY estimates, and the six ICERs for pairwise comparisons of the therapies also increase in magnitude by up to 100%.

These changes are potentially large enough to alter policy decisions. For example, the ICER comparing monotherapy with combination DMARD+steroid rises from £13,721 to £21,455 using the independent domains model. If we were to use a cost-effectiveness threshold of £20,000, this would reverse the decision that monotherapy is cost-effective relative to the DMARD+steroid combination therapy. Using the joint model, the ICER rises to £18,100, not large enough to reverse the decision but a substantial rise nonetheless.

So there is some support here for the practical use of the independent domains model (which is implemented in the Stata command `bicop`) as the predicted QALYs are similar to the joint model. In this specific case, those small differences still translate into substantial

differences in the cost-effectiveness estimates because the ICERs are very sensitive to the incremental QALYs.

Table 7: Mean costs, QALYs and incremental cost-effectiveness ratios for the CARDERA trial

	Monotherapy	Combination therapies		
	MTX	MTX+CS	MTX+PNS	MTX+CS+PNS
Total costs[1]	£7,503	£6,829	£6,323	£6,203
EQ-5D-3L from trial data				
Total QALYs	1.238	1.093	1.152	1.320
ICER (for col therapy vs. row therapy)				
MTX only	-	£4,648	£13,714	-£15,929
MTX+CS	£4,648	-	-£8,597	-£2,765
MTX+PNS	£13,714	-£8,597	-	-£714
EQ-5D-5L mapped from 3L trial data (independent domains model)				
Total QALYs	1.452	1.368	1.397	1.523
ICER (for col therapy vs. row therapy)				
MTX only	-	£8,021	£21,476	-£18,254
MTX+CS	£8,021	-	-£17,440	-£4,037
MTX+PNS	£21,476	-£17,440	-	-£952
EQ-5D-5L mapped from 3L trial data (joint model)				
Total QALYs	1.440	1.343	1.375	1.504
ICER (for col therapy vs. row therapy)				
MTX only	-	£6,930	£18,100	-£20,141
MTX+CS	£6,930	-	-£15,819	-£3,873
MTX+PNS	£18,100	-£15,819	-	-£926

[1] Present value of treatment costs over the 2-year experimental period

7 Conclusions

There are three clear conclusions. First, econometric modelling based on a flexible mixture-copula specification has revealed significant differences between the 3-level and 5-level versions of the EQ-5D descriptive system for health states. These differences are particularly striking for the mobility and pain domains, where the two versions of the instrument give significantly different pictures of the relationship between individual health states and their demographic and clinical determinants.

Second, we have developed a new and powerful technique for modelling and mapping between the 3-level and 5-level versions of EQ-5D, using an Empirical Bayes conditional expectation approach. This has revealed some asymmetry. Mapping from EQ-5D-3L questionnaire responses to predictions of the utility scores for EQ-5D-5L reproduces the directly-observed distributional shape quite faithfully. In contrast, mapping from the more detailed EQ-5D-5L responses to predicted utility scores for EQ-5D-3L fails to capture the bunching of utilities at low or negative values that is characteristics of EQ-5D-3L. This has important implications for policy bodies like NICE. On the basis of the evidence presented here, NICE could move to the new 5-level version of EQ-5D as the basis for its decision-making, and use flexible mapping techniques where necessary to convert old 3-level EQ-5D evidence to the new basis. It would be unwise to convert newer 5L-based trial evidence back to the old 3L basis, since mapping does not work well in that context.

Third, our re-examination of evidence from a trial of combination drug therapies for rheumatoid arthritis shows that switching to the newer 5-level version of EQ-5D can make a substantial difference to the conclusions from cost-effectiveness studies, so there is likely to be a need to re-examine past decisions to investigate their robustness. Our new mapping approach offers a way of doing this and is readily applied using recently-developed computer code implemented as a Stata command `bicop` (Hernández-Alava and Pudney, 2016). The mapping algorithms developed in this paper will also be made freely available as a Stata command for analysts to use in RA. The threshold historically used for EQ-5D-3L may require reassessment if the 5-level version of EQ-5D is to be used in future.

References

Agborsangaya, C. B., Lahtinen, M., Cooke, T., and Johnson, J. A. (2014). Comparing the EQ-5D 3L and 5L: measurement properties and association with chronic conditions and multimorbidity in the general population. *Health and Quality of Life Outcomes*, 12:1–7.

Augustovski, F., Rey-Ares, L., Irazola, V., Garay, O. U., Gianneo, O., Fernández, G., Morales, M., Gibbons, L., and Ramos-Goñi, J. M. (2015). An eq-5d-5l value set based on uruguayan population preferences. *Quality of Life Research*, pages 1–11.

Bedford, T. and Cooke, R. (2002). Vines - a new graphical model for dependent random variables. *The Annals of Statistics*, 30:1031–1068.

Brazier, J. E., Yang, Y., Tsuchiya, A., and Rowen, D. L. (2010). A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *The European Journal of Health Economics*, 11(2):215–225.

Bruce, B. and Fries, J. F. (2003). The Stanford Health Assessment Questionnaire (HAQ): a review of its history, issues, progress, and documentation. *Journal of Rheumatology*, 30:67–78.

Choy, E. H. S., Smith, C. M., Farewell, V., Walker, D., Hassell, A., Chau, L., and Scott, D. L. (2008). Factorial randomised controlled trial of glucocorticoids and combination disease modifying drugs in early rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 67:656–663.

Claxton, K., Martin, S., Rice, N., Spackman, E., Hinde, S., Devlin, N., Smith, P. C., and Sculpher, M. (2015). Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. *Health Technology Assessment*, 19(14).

Devlin, N., Shah, K., Feng, Y., Mulhern, B., and van Hout, B. (2016). Valuing health-related quality of life: An EQ-5D-5L value set for England. Technical Report 16.02, Health Economics & Decision Science, University of Sheffield.

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care*, 35:1095–1108.

Fayers, P. M. and Hays, R. D. (2014). Should linking replace regression when mapping from profile-based measures to preference-based measures? *Value in Health*, 17(2):261 – 265.

Feng, Y., Devlin, N., Shah, K., Mulhern, B., and van Hout, B. (2016). New methods for modelling EQ-5D-5L value sets: an application to English data. Technical Report 16.03.

Hernández-Alava, M. and Pudney, S. E. (2016). BICOP: A command for estimating bivariate ordinal regressions with residual dependence characterized by a copula function and normal mixture marginal. *Stata Journal*, forthcoming.

Hernández-Alava, M., Wailoo, A. J., and Ara, R. (2012). Tails from the peak district: Adjusted limited dependent variable mixture models of EQ-5D health state utility values. *Value in Health*, 15:550–561.

Ikeda, S., Shiroiwa, T., IgarashiA., NotoS., FukudaT., SaitoS., and Shimozuma, K. (2015). Developing a japanese version of the eq-5d-5l value set. *Journal of the National Institute of Public Health*, 64(1):47–55.

Janssen, M. F., Birnie, E., and Bonsel, G. J. (2008a). Quantification of the level descriptors for the standard EQ-5D three-level system and a five-level version according to two methods. *Quality of Life Research*, 17:463–473.

Janssen, M. F., Birnie, E., Haagsma, J. A., and Bonsel, G. J. (2008b). Comparing the standard EQ-5D three-level system with a five-level version. *Value in Health*, 11:275–284.

Janssen, M. F., Pickard, A. S., Golicki, D., Gudex, C., Niewada, M., Scalone, L., Swinburn, P., and Busschbach, J. (2013). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Quality of Life Research*, 22:1717–1727.

Jia, Y. X., Cui, F. Q., Li, L., Zhang, D. L., Zhang, G. M., Wang, F. Z., Gong, X. H., Zheng, H., Wu, Z. H., Miao, N., Sun, X. J., Zhang, L., Lv, J. J., and Yang, F. (2014). Comparison between the EQ-5D-5L and the EQ-5D-3L in patients with hepatitis B. *Quality of Life Research*, 23:2355–2363.

Longworth, L. and Rowen, D. (2011). Nice dsu technical support document 10: The use of mapping methods to estimate health state utility values.

NICE (2013). Guide to the methods of technology appraisal 2013. Technical report, National Institute for Health and Care Excellence.

NICE, editor (2014). *Developing NICE guidelines: the manual*.

Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107:1063–1072.

Pattanaphesaj, J. and Thavorncharoensap, M. (2015). Measurement properties of the EQ-5D-5L, compared to EQ-5D-3L in the Thai diabetes patients. *Health and Quality of Life Outcomes*, 13:1–8.

Pickard, A. S., Leon, M. C. D., Kohlmann, T., Cella, D., and Rosenbloom, S. (2007). Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. *Medical Care*, 45:259–263.

Rivero-Arias, O., Ouellet, M., Gray, A., Wolstenholme, J., Rothwell, P. M., and Luengo-Fernandez, R. (2010). Mapping the modified rankin scale (mrs) measurement into the generic euroqol (eq-5d) health outcome. *Medical Decision Making*, 30(3):341–354.

Scalone, L., Ciampichini, R., Fagiuoli, S., Gardini, I., Fusco, F., Gaeta, L., Prete, A. D., Cesana, G., and Mantovani, L. G. (2013). Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. *Quality of Life Research*, 22:1707–1716.

Trivedi, P. K. and Zimmer, D. M. (2005). Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1:1–111.

Wailoo, A., Hernández-Alava, M., Scott, I. C., Ibrahim, F., and Scott, D. L. (2014). Cost-effectiveness of treatment strategies using combination disease-modifying anti-rheumatic drugs and glucocorticoids in early rheumatoid arthritis. *Rheumatology*, 53:1773–1777.

Wolfe, F. and Michaud, K. (2011). The National Data Bank for rheumatic diseases: a multi-registry rheumatic disease data bank. *Rheumatology*, 50:16–24.

Xie, F., Pullenayegum, E., Gaebel, K., Bansback, N., Bryan, S., Ohinmaa, A., Poissant, L., and Johnson, J. A. (2016). A time trade-off-derived value set of the eq-5d-5l for canada. *Medical Care*, 54(1):98–105.

Appendix: full parameter estimates

Table A1 Estimated coefficients of the domain-specific bivariate and joint models

	Domain-specific model		Joint model	
	Coefficient	Std. error	Coefficient	Std. error
Mobility domain - 3 levels				
male	0.4601	0.0543	0.5125	0.0637
age/10	-0.0117	0.0169	-0.0067	0.0197
pain/10	2.4178	0.3205	2.8928	0.3826
HAQ	1.2370	0.1092	1.3765	0.1347
HAQ^2	-0.9591	0.3880	0.0987	0.0627
$pain^2$	0.0593	0.0522	-1.2067	0.4554
HAQ * pain	-0.3067	0.1603	-0.3134	0.1907
ψ			0.6494	0.0416
Γ_1	1.8996	0.1244	2.2583	0.1547
Γ_2	5.6557	0.1634	6.7752	0.2465
Mobility domain - 5 levels				
male	0.3390	0.0430	0.3839	0.0504
age/10	0.0506	0.0137	0.0612	0.0159
pain/10	1.9446	0.2525	2.4359	0.2964
HAQ	1.2235	0.0841	1.4009	0.1010
HAQ^2	-0.4122	0.3099	0.0610	0.0470
$pain^2$	0.0458	0.0397	-0.6556	0.3606
HAQ * pain	-0.3969	0.1283	-0.4656	0.1527
ψ			0.6279	0.0317
Γ_1	1.5939	0.0982	1.8964	0.1184
Γ_2	2.9367	0.1032	3.4302	0.1321
Γ_3	4.2711	0.1093	4.9911	0.1511
Γ_4	5.5625	0.1303	6.5589	0.1920
Dependency θ	0.7074	0.0139	0.5956	0.0203

continued...

Table A1 continued

	Domain-specific model		Joint model	
	Coefficient	Std. error	Coefficient	Std. error
Self-care domain - 3 levels				
male	0.6103	0.0662	0.6438	0.0688
age/10	-0.1067	0.0204	-0.1096	0.0210
pain/10	1.0591	0.4462	1.4948	0.4722
HAQ	1.8555	0.1966	1.9641	0.2226
HAQ^2	-0.6821	0.4457	-0.0444	0.0790
$pain^2$	-0.0314	0.0729	-1.0048	0.4603
HAQ * pain	0.0428	0.2036	0.0040	0.2144
ψ			0.3163	0.0347
Γ_1	2.7358	0.1960	2.9350	0.2235
Γ_2	5.7598	0.2142	6.1590	0.2565
Self-care domain - 5 levels				
male	0.6366	0.0536	0.6779	0.0569
age/10	-0.0949	0.0167	-0.1006	0.0175
pain/10	1.2139	0.3390	1.7335	0.3669
HAQ	1.5870	0.1270	1.7245	0.1432
HAQ^2	-0.7787	0.3644	0.0097	0.0561
$pain^2$	0.0182	0.0519	-1.1726	0.3852
HAQ * pain	0.0764	0.1583	0.0276	0.1686
ψ			0.3806	0.0289
Γ_1	2.0816	0.1350	2.3131	0.1524
Γ_2	3.4855	0.1399	3.7768	0.1627
Γ_3	4.9402	0.1512	5.3745	0.1825
Γ_4	5.6903	0.1729	6.3115	0.2176
Dependency θ	6.0530	0.3145	5.5022	0.3051

continued...

Table A1 continued

	Domain-specific model		Joint model	
	Coefficient	Std. error	Coefficient	Std. error
Usual activities domain - 3 levels				
male	0.2409	0.0539	0.3278	0.0781
age/10	-0.0582	0.0168	-0.0751	0.0240
pain/10	2.6254	0.3175	4.1937	0.4879
HAQ	1.7515	0.1164	2.6488	0.1936
HAQ^2	-1.3382	0.3756	-0.3058	0.0709
$pain^2$	-0.1891	0.0503	-2.1676	0.5438
HAQ * pain	0.0196	0.1594	-0.1170	0.2237
ψ			1.0333	0.0819
Γ_1	1.7532	0.1278	2.7194	0.2159
Γ_2	4.7465	0.1520	6.9414	0.3559
Usual activities domain - 5 levels				
male	0.1923	0.0440	0.2462	0.0625
age/10	-0.0751	0.0139	-0.0961	0.0195
pain/10	2.4151	0.2616	3.7146	0.3862
HAQ	1.6059	0.0925	2.2971	0.1437
HAQ^2	-1.3418	0.3149	-0.1997	0.0581
$pain^2$	-0.1386	0.0416	-2.0802	0.4497
HAQ * pain	0.0367	0.1325	-0.0395	0.1881
ψ			0.9943	0.0616
Γ_1	1.0144	0.0997	1.5766	0.1490
Γ_2	2.4708	0.1074	3.6049	0.1854
Γ_3	3.9116	0.1188	5.6372	0.2345
Γ_4	4.8488	0.1342	6.8882	0.2712
Dependency θ	0.5560	0.0172	0.1019	0.0541
Common mixture				
π	0.0621	0.0461		
$1-\pi$	0.9379	0.0461		
μ_1	0.2841	0.4314		
μ_2	-0.0188	0.0217		
σ_1^2	3.0482	0.8537		
σ_2^2	0.8587	0.0665		

continued...

Table A1 continued

	Domain-specific model		Joint model	
	Coefficient	Std. error	Coefficient	Std. error
Pain/discomfort domain - 3 levels				
male	0.1737	0.0472	0.2130	0.0562
age/10	0.0332	0.0156	0.0274	0.0181
pain/10	6.3976	0.4445	7.1520	0.4037
HAQ	0.6059	0.0908	0.7806	0.1046
HAQ^2	-2.3849	0.4493	-0.1176	0.0551
$pain^2$	-0.1296	0.0488	-3.0418	0.4349
HAQ * pain	0.4015	0.1796	0.1717	0.1849
ψ			0.3705	0.0325
Γ_1	0.8379	0.1132	0.9465	0.1241
Γ_2	5.1633	0.1728	5.4769	0.1890
π	0.5871	0.0787		
$1-\pi$	0.4129	0.0787		
μ_1	-0.0936	0.0528		
μ_2	0.1331	0.0771		
σ_1^2	0.2850	0.0824		
σ_2^2	1.9866	0.2359		
Pain/discomfort domain - 5 levels				
male	0.1085	0.0424	0.1278	0.0484
age/10	-0.0504	0.0137	-0.0605	0.0155
pain/10	6.0189	0.2887	6.9250	0.3362
HAQ	0.6694	0.0819	0.7903	0.0936
HAQ^2	-2.6218	0.3451	-0.1119	0.0460
$pain^2$	-0.1042	0.0402	-3.0565	0.3848
HAQ * pain	0.3632	0.1391	0.3352	0.1563
ψ			0.5364	0.0301
Γ_1	-0.3351	0.0939	-0.3981	0.1061
Γ_2	2.0121	0.1049	2.3200	0.1212
Γ_3	4.1984	0.1174	4.7505	0.1437
Γ_4	5.3824	0.1280	6.0899	0.1616
π	0.1075	0.0745		
$1-\pi$	0.8925	0.0745		
μ_1	0.1204	0.1985		
μ_2	-0.0145	0.0195		
σ_1^2	2.6886	0.7068		
σ_2^2	0.7948	0.0830		
Dependency θ	1.7094	0.0474	1.5660	0.0452

continued...

Table A1 continued

	Domain-specific model		Joint model	
	Coefficient	Std. error	Coefficient	Std. error
Anxiety/depression domain - 3 levels				
male	0.0387	0.0491	0.0469	0.0495
age/10	-0.1350	0.0148	-0.1355	0.0152
pain/10	1.2087	0.2829	1.3453	0.2894
HAQ	0.4322	0.0904	0.4549	0.0923
HAQ^2	-0.2623	0.3495	-0.0663	0.0440
$pain^2$	-0.0580	0.0436	-0.4026	0.3550
HAQ * pain	0.1788	0.1471	0.1903	0.1478
ψ			0.3257	0.0259
Γ_1	0.4435	0.1033	0.4901	0.1055
Γ_2	2.2668	0.1086	2.3920	0.1164
Anxiety/depression domain - 5 levels				
male	-0.0137	0.0453	-0.0071	0.0462
age/10	-0.1456	0.0137	-0.1482	0.0142
pain/10	1.2094	0.2554	1.3614	0.2640
HAQ	0.3731	0.0826	0.4139	0.0855
HAQ^2	-0.4111	0.3179	-0.0526	0.0410
$pain^2$	-0.0387	0.0401	-0.5557	0.3251
HAQ * pain	0.2730	0.1354	0.2818	0.1377
ψ			0.3554	0.0240
Γ_1	0.1154	0.0945	0.1625	0.0979
Γ_2	1.0888	0.0953	1.1589	0.0999
Γ_3	2.0811	0.0998	2.2051	0.1076
Γ_4	2.6195	0.1098	2.8087	0.1227
Dependency θ	14.4849	0.5894	13.9413	0.5912
Common mixture - Joint model				
π			0.0250	0.0127
$1 - \pi$			0.9750	0.0127
μ_1			-0.5004	0.2528
μ_2			0.0128	0.0072
σ_1^2			5.6660	1.6944
σ_2^2			0.8739	0.0286