DEVELOPMENT OF AN ALGORITHM TO DERIVE 'UTILITIES' FROM A VALIDATED MENOPAUSE-SPECIFIC HRQOL QUESTIONNAIRE

Zöllner YF[1], Brazier JE[2], Oliver P[2], Alt J[3], Schaefer M[1]

[1]Dept. of Social Pharmacy, Humboldt University, Berlin, Germany;
[2]School of Health and Related Research, University of Sheffield, UK
[3]Women's Health Unit, Solvay Pharmaceuticals, Hannover, Germany

ABSTRACT

Background Quality-adjusted life-years (QALYs) are the recommended metric for assessing the economic value of many preventive and long-term treatments, such as hormone replacement therapy (HRT). QALYs combine quality of life (QoL) and life expectancy in a single construct, whereby QoL is measured in 'utils' on a preference scale anchored on '1' for 'perfect health and '0' for 'death'. In the climacteric field, most cost-utility analyses (CUAs) found in the literature rely on arbitrarily assumed rather than empirically derived utility weights. Some attempts to derive preference scores for menopausal health states have been undertaken but have so far only addressed isolated events within the greater menopause framework. **Objective** To develop an algorithm to convert menopausal health profiles – generated through a previously validated questionnaire – into preference-based QoL weights. **Methods** A newly developed menopause-specific scale, the "QualiPause Inventory" (QPI), has been tested for internal consistency, retest reliability, face-, content- and construct validity. Multiple regression models of SF-36 and EQ-5D-derived utility scores against QPI sub-scale scores have been run in order to determine an 'exchange rate' between menopausal QoL domain scores and overall utilities. A compact 6-dimensional health state classification system has been derived from the QPI, from which a sample of 15-20 health states will be selected for direct valuation using the time trade-off (TTO) technique. Regression models will be used to inter- and extrapolate the utility weights of the remaining (not directly valued) health states. **Results** The QPI fulfils the set criteria for validity and reliability. Regression shows that the more generic QPI domains (psychosocial, physical) are good predictors of SF-36 and EQ-5D-derived QoL scores. However, the condition- and treatment-specific domains (vasomotor, menstrual, sexual, androgenic) are poorly related to the latter. **Conclusion** Multi-attribute utility scales can provide useful but limited QoL scores of postmenopausal health states. A direct valuation exercise in necessary to reveal the importance attached to the menopause-specific QoL domains. An algorithm able to convert menopausal health profiles into overall utility weights would be highly desirable.

BACKGROUND

The usefulness of cost-utility analyses for decision-making in health care critically depends on the existence of meaningful QALY estimates for the health states involved in a particular condition and its treatment; this is especially true in the field of the climacteric and the menopause, i.e. the health assessment of mid-aged and elderly women. Economic evaluations of postmenopausal treatments, and hormone replacement therapy (HRT) in particular, still lack appropriate health-related quality-of-life (QoL) weights attached to postmenopausal health states. This is apparent from the discrepancy between utility values given to similar health states found in the literature. Furthermore, there is no consensus on how to derive such utilities. QoL weights have, in this area, traditionally been arbitrarily assumed. Following the seminal example of Daly *et al*. (1993), only a few recent attempts have been made to elicit these empirically (Zethraeus *et al*. 1997; Gabriel *et al*. 1999; Tosteson *et al*. 2000). Results, though not contradictory, vary considerably by study design and elicitation method. No study thus far includes an integrated and comprehensive evaluation of both menopausal symptom relief and HRT side-effects; often, only isolated menopausal symptoms are covered, or the description of health states is fairly vague. Utility (and 'disutility') arising from the prospects of long-term prevention and/or potential complications have been addressed in part, and might represent yet another challenging study objective, although probably beyond the scope of this particular project.

PRELIMINARY WORK

From an exhaustive literature review of climacteric health and QoL instruments, no scale including the full range of potential side-effects of HRT could be identified. Therefore, and as a foundation for further research into 'utilities', a comprehensive patient-reported health questionnaire — covering both menopausal symptoms and potential treatment side-effects — has been developed and validated (Zöllner *et al*., 2001). The final instrument ("QualiPause Inventory", QPI) includes 22 items which are measured on a 5-point 'botheredness' scale[1] and spread across 6 domains, explaining

[1] Scale points are 'absent', followed by 'hardly, moderately, considerably' and 'extremely' bothered, respectively.

62% of data variance. The 6 sub-scales are psychosocial (PSY), vasomotor (VASO), physical (PHYS), sexual (SEX), menstrual (MENS), and androgenic (ANDRO). The key characteristics of the questionnaire are summarised in **table 1**.

Table 1. Key psychometric data of the "QualiPause Inventory" (QPI).

item label	retest *r*	discriminant[1] *p*	factor loading	domain label	Cronbach's alpha	% variance explained
Difficulty sleeping	.73	***	.45	**Psycho-social (PSY)**	0.90	28.7
Feeling tense or nervous	.67	***	.89			
Feeling restless or irritable	.66	***	.78			
Being easily worried	.71	***	.86			
Anxious or frightened feelings	.71	**	.92			
Panic attacks	.74	***	.69			
Mood swings	.54	***	.64			
Tiredness or fatigue	.70	***	.44			
Hot flushes	.79	** (users *vs.* non-users)	.88	**Vasomotor (VASO)**	0.82	8.9
Night sweats	.68	*** (sympt. *vs.* non-sympt.) * (users *vs.* non-users)	.88			
Stiffness of joints	.64	***	.84	**Physical (PHYS)**	0.79	7.6
Swelling of limbs	.59	***	.57			
Back ache	.63	***	.73			
Aching joints or muscles	.61	***	.90			
Vaginal dryness	.40	**	.88	**Sexual (SEX)**	0.74	6.7
Pain at intercourse	.57	p=.055	.89			
Monthly bleeding	.72	NS (2)	.81	**Menstrual (MENS)**	0.5	5.4
Irregular bleeding	.73	NS (2)	.81			
Breast tenderness	.55	***	.48			
Loss of hair from scalp	.66	p=0.053 (2)	.42	**Androgenic (ANDRO)**	0.4	5.3
Greasy skin or acne	.71	NS (2)	.81			
Facial or body hair growth	.71	NS (2)	.73			

[1] Discrimination between "symptomatic" and "non-symptomatic" (based on lead symptom: hot flushes) unless otherwise stated.
[2] Discrimination between HRT users and non-users as these are potential adverse effects

The novelty to this questionnaire is that it covers not only all signs and symptoms relevant and specific to the menopause and the climacteric; it also includes those signs and symptoms potentially arising after onset of HRT, such as androgen- and oestrogen-borne side-effects (see the androgenic domain for the former, and the menstrual domain for the latter). Irregular bleeding and regular bleeding, apart from occurring naturally, can be HRT-induced and are then (fairly) mutually exclusive. Regular bleeds occur with the traditional sequential (=cyclical) regimens, irregular bleeds can occur with the newer continuous dosage forms. Bleeding pattern, controversially discussed in terms of its QoL impact at present, has not been differentially captured thus far by any standardised questionnaire, either.

The new questionnaire fulfils the requirements for stability (test-retest reliability) and item convergent and item divergent validity. It also shows very good discriminant or known-groups validity. The ability of the "bleeding" and "androgenic" items to discriminate between users and non-users of HRT, though not apparent here, may show under more controlled conditions (clinical trial). Analogously, the somewhat unexpectedly low Cronbach's alpha value for the androgenic domain might be specific to this particular sample; in either case, there is a strong medical rationale for leaving in those items for the benefit of a comprehensive, in-depth assessment of the well-being of climacteric and post-menopausal women. The item candidate ‘Being bothered by body weight’, also potentially associated with HRT use, has eventually *not* been included in the questionnaire itself as weight can be readily and objectively quantified (in kg or pound/stone) and therefore be captured in the preceding screening section.

Regarding concurrent validity, the QPI's psychosocial and physical sub-scales showed very good correlations with the pertinent sub-scales of both the SF-36 and the EQ-5D (**table 2**).

The instrument can be used off-the-shelf in clinical trials and observational studies. Especially to the clinician, who is interested in particular symptoms (or groups of symptoms) and the extent to which these respond to the treatment prescribed, this should prove useful. The health policy maker, in turn, will rather seek an overall index and, more to the point, a *preference-based* index. The latter is fundamentally different from the questionnaire ‘overall index’ which aggregates all items with uniform weights.

A preference-based index will summarise any health state profile into a single number between 0 and 1, where 0 equals 'death' and 1 equals 'full health' on an interval (or ratio) scale.

Table 2. Pearson correlation coefficients between psychological/somatic domains and the SF-36 and EQ-5D. Expected significant correlations in bold.

		QPI sub-scale	
		Psychosocial	Physical
SF-36 sub-scale	Mental Health Index	**0.76**	0.35
	Role-Emotional	**0.62**	0.33
	Energy-Vitality	**0.62**	0.50
	Energy-Vitality	**0.62**	0.50
	Role-Functional	0.48	**0.50**
	Pain Index	0.42	**0.73**
	General Health Perceptions	0.45	**0.52**
	Physical Functioning	0.29	**0.58**
EQ-5D sub-scale	Psychological	**0.66**	0.25
	Mobility	0.19	**0.46**
	Activities	0.36	**0.50**
	Pain	0.27	**0.62**

OBJECTIVE

Being a health economics project, the aim of this piece of research is to match up every possible menopausal health state or profile (response pattern to the QPI), with a *preference-based single index*, or *utility score* (*syn*. QoL weight, preference weight).

There are essentially three approaches for estimating a preference-based single index measure for the QPI. One is to conduct a genuine valuation survey using preference-based techniques such as the standard gamble (SG) or time trade-off (TTO). A second alternative is a mapping exercise where the responses to the questionnaire under investigation are mapped onto a generic utility-generating instrument (a multi-attribute utility scale) such as the EQ-5D. The third approach consists of deriving 'exchange rates' between the QPI and an existing preference-based measure.

The second approach has been discarded for reasons of non-feasibility (impossible to map some items very closely related to menopause onto general health items) and uncertainty (even with explicitly derived mapping rules, no evidence on their predictive validity).

The focus of this paper is on the first approach, i.e. a genuine valuation survey, the design of which is currently under development. Some light will be shed on the third approach, i.e. 'exchange rate modelling', in the exploratory part of the 'methods' section.

The goal of matching health profiles to utilities has been achieved e.g. for the EQ-5D, where a 'social tariff' could be derived for all of the $3^5 = 243$ possible health states. In the case of the QPI, many millions of health states are defined by the questionnaire. This situation requires a previous step, i.e. the development of a simplified health state classification system (HSCS). Given the hexa-dimensionality of the QPI, the HSCS should reflect all the six dimensions of the QPI. The work by Brazier *et al.* (1998) is conceptually similar in that it first transforms the SF-36 into the six-dimensional SF-6D and then elicits utilities corresponding to the latter. This project intends to reproduce this approach with a condition-specific instrument, namely the QPI.

Eventually, either the HSCS questionnaire or the full-length inventory could be administered in a clinical trial or observational study. In the case of administering the full version, the resulting response pattern could directly and unambiguously be mapped onto the HSCS which in turn would yield the utility scores.

METHODS

1. Exploratory work

In order to get a first feel for the work with external single indices and how they relate to the QPI dimensions, some correlations and regressions have been carried out. The exploratory work outlined here would represent the 'exchange rate' approach between a generic preference-based index and a condition-specific health profile as mentioned above.

The dataset generated during the QPI validation phase included the menopausal questionnaire itself, as well as the SF-36 (UK) and the EQ-5D, all of which were completed by 785 women (post data-cleaning *n*). Each case (woman) was assigned

two utility values, one stemming from the SF-36/SF-6D work by Brazier *et al*. (*op.cit.*), and one stemming from the EQ-5D (Dolan *et al*., 1997).

The SF-6D and EQ-5D-derived utility values (U) (which were reasonably consistent among each other, r =0.65) were analysed in two ways: (1) correlation with QPI domain (or sub-scale) scores, and (2) multiple regression using the QPI domain scores as predictor variables. The results for the correlation are shown in **table 3**.

Table 3. Pearson correlation coefficients between the QPI sub-scale scores and the SF-6D and EQ-5D-derived utilities, respectively.

	PSY	VASO	PHYS	SEX	MENS	ANDRO
SF-6D index	.664***	.302***	.599***	.150**	.165***	.233***
EQ-5D index	.452***	.241***	.559***	.106[a]	.071[b]	.216***

*** $p < .0001$, ** $p < .001$, [a] $p = 0.021$, [b] $p = 0.051$

Multiple regression analysis was carried out to assess how well the utility values (U) from the SF-6D and EQ-5D could be predicted by the 6 QPI domains. The model is defined as:

$$\mathbf{U} = ß_0 + ß_1 PSY + ß_2 VASO + ß_3 PHYS + ß_4 SEX + ß_5 MENS + ß_6 ANDRO \qquad \textbf{[Eq. 1]}$$

where U = utility (SF-6D or EQ-5D-derived) and the variables (in capital letters) represent the sub-scales of the QPI. The Beta coefficients (ß) in **equation 1** give an idea of the relative weight attached to a particular group of symptoms, thus providing an avenue to future 'exchange rate' modelling between the SF-6D utilities and the QPI.

Table 4 shows the results of the multiple regression analysis for both models.

Table 4. Model summaries corresponding to eq. 1.

		Model				Individual Coefficients		
		Summary		ANOVA		beta	t	Sig
				F	Sig			
Dependent Variables	Predictor Variables	R	R^2					
SF-6D-derived utility index		**0.747**	**0.558**	**92.53**	**<0.001**			
	PSY					-8.24×10^{-3}	13.17	***
	PHYS					-1.29×10^{-2}	10.33	***
	SEX					6.12×10^{-3}	2.87	**
	MENS					-1.57×10^{-3}	0.68	
	VASO					1.16×10^{-3}	0.53	
	ANDRO					-1.03×10^{-3}	0.41	
EQ-5D-derived utility index		**0.607**	**0.369**	**45.08**	**<0.001**			
	PHYS					-2.78×10^{-2}	10.45	***
	PSY					-8.12×10^{-3}	5.93	***
	SEX					1.05×10^{-2}	2.31	*
	ANDRO					-6.06×10^{-3}	1.11	
	MENS					4.43×10^{-3}	0.86	
	VASO					1.53×10^{-3}	0.33	

*** $p<0.001$; ** $p<0.01$; * $p<0.05$

The amount of $U_{SF\text{-}6D}$ variance explained by the model (eq. 1) was 56% and an analysis of variance (ANOVA) was significant ($F=92.53$, $p<0.001$). We therefore conclude that this model fitted the observed data well. Three of the predictor variables (PSY, PHYS) contribute significantly to the regression equation; however, the MENS, VASO and ANDRO domains do not. The SEX domain is out of line with the expected pattern as it does contribute significantly to the regression but with the wrong sign.

The overall amount of $U_{EQ\text{-}5D}$ variance explained by this model was less than for the SF-6D at 37%; however the ANOVA for this model was significant ($F=92.53$, $p<0.001$). Again, the predictor variables PSY and PHYS contributed significantly to the regression whereas MENS, VASO and ANDRO did not; SEX did at a rather low significance level, and still with the wrong sign.

Focussing again on the *direct estimation* of preference weights from the QPI – the objective of this study – we find that this involves three steps: first; to derive a HSCS from the QPI that is amenable to valuation; second, to conduct a valuation survey of a sample of states defined by the classification; and third, to build a regression model to estimate the remaining weights.

2. Development of the health state classification system (HSCS)

In order to represent the QPI in full, and as mentioned above, every one of the six dimensions should be represented. Six items of QPI – one per domain – have been earmarked as 'best representatives' of their respective domains, based on factor loadings (item-convergent and item-divergent validity), discriminant or known-groups validity, test-retest reliability, and 'importance' (the product of frequency by severity), and are shown in **table 5**. These key items shall serve as a basis for the construction of a HSCS, to be called "QPI-6D" hereafter.

Table 5. The six items/dimensions for the "QPI-6D".

domain	item	scaling concept	possible combination of items
psychosocial	"anxious or frightened feelings"	frequency	yes: 'panic attacks' as worst level of 'anxiousness/ frightened feelings'
vasomotor	"hot flushes"	frequency	yes: 'hot flushes' and 'night sweats' together ("or"-connected) given close correlation
physical	"aching joints and muscles"	frequency, intensity, or both (?)	no (covers ± everything)
sexual	"pain at intercourse"	frequency	yes: 'vaginal dryness' and 'pain at intercourse' given close correlation
menstrual	"breast tenderness"	intensity	difficult to combine, though bleeding question could be interesting to address
androgenic	"greasy skin or acne"	intensity	no

Expressing each item (concept) on a basis of five levels (one for 'absence' of symptom and four representing frequency or severity), the QPI-6D would define 5^6 = 15,625 distinct health states. If one 'compressed' the scale (obviously sacrificing some sensitivity) by lumping together the two lower and two upper scale points for all dimensions, there would only be 3^6 = 729 possible health states. A possible HSCS framework is outlined in **table 6**.

Table 6. Simple framework for the QPI-6D HSCS.
This structure would define $5^2 * 3^4 = 25 * 81 = 2{,}025$ health states.

domain	scaled items
psychosocial	you do not feel anxious or frightened. you rarely feel anxious or frightened you sometimes feel anxious or frightened. you often feel anxious or frightened. you often get panic attacks.
vasomotor	you do not get any hot flushes or night sweats. you rarely get hot flushes or night sweats you occasionally get hot flushes or night sweats. you frequently get hot flushes or night sweats. You continuously have hot flushes or night sweats.
physical	you do not have aching joints or muscles. you have mild to moderate joint or muscle aches. you have considerable o severe joint or muscle aches.
Sexual (if sexually active)	you do not suffer from vaginal dryness or pain at intercourse. you sometimes suffer from vaginal dryness/pain at intercourse. you often suffer from vaginal dryness/pain at intercourse.
menstrual	you do not have any breast tenderness. you have mild to moderate breast tenderness. you have considerable to severe breast tenderness.
androgenic	you do not have greasy skin or acne. you have mild to moderate greasy skin or acne. you have considerable to severe greasy skin or acne.

4. Valuation survey (planned work, at design stage)

Scientific part

Even with 'only' 2,025 possible health states, a manageable 'survey' sample of these will have to be selected for valuation; the preference values of the remaining health states having to be estimated by inter- and extrapolation. Orthoplan, a feature embedded in the SPSS package, will generate such a sample of health states. Based on initially 2,025 health states, a sample of approximately 15-20 health states would typically be generated, all of which must obey to a certain set of rules:

- Every health state description will be 'complete', i.e. encompass all six domains;
- Every level of each dimension should appear in at least one health state description;
- States should include different combinations of dimension levels, and in such a way that the number of non-axiomatic choices to be made is maximised (or the number of clearly dominant alternatives minimised)
- Only 'real' health states, i.e. those actually being observed in the existing datasets should be valued.

In order to provide a reasonably robust model, as well as for testing purposes, we will attempt to value more health states than those strictly generated by orthoplan.

There are no clear-cut rules to sample size calculation in this sort of exercise. Assuming a statistical requirement of about 10-12 valuations (observations) per health state, this would amount to about 250 individual valuations. Every women cannot be expected to value – consistently and reliably – any more than 8 health states; one of these should be her own current health state (which we cannot influence), thus leaving 7 'imposed' ones for valuation. Allowing for some inconsistencies, missing answers etc, we should hence expect to survey approximately 50 to 60 women.

The reason for asking women to value their own current health state as part of the exercise is to work as close to reality as possible. In the case of inconsistencies among answers, the 'real life' health state can probably be given greater faith. Also, a concomitant administration of the SF-36 and/or EQ-5D which women would complete to reflect their own current health status would generate additional utility values which can be correlated to the directly elicited utilities or to test the predictive ability of any proposed regression models. The TTO has been decided to be the valuation method of

choice for this study; however, we are also looking into a concomitant administration of a conjoint-based test procedure.

Prior to embarking on the valuation survey, a small 'pilot' should be launched, i.e. a "dummy run" of the devised TTO (and CA) instrument to check for understandability and clarity especially of the method. This would take place in in-depth, face-to-face interviews with a small number of respondents (say, 5, who could even be university staff). Issues arising during this phase can then still feed back into the instrument design and/or instruction of the interviewer/instrument administrator.

Logistic part

694 women (99.5% in the HRT user group, 89.5% in the non-user group) asked for feedback as part of the original validation exercise (postal survey) in December 2000/January 2001. Along with the report they will be sent, they'll be asked whether they would be willing to take part in a follow-up activity (reply-paid post card, including woman's address and telephone). Those showing interest will be invited (by phone; propose three alternative dates) to a 'peer meeting' on hospital premises in groups of about 10. Five or six of these meetings would then need to be scheduled, all before the end of this year.

DISCUSSION

In terms of the 'exchange rate' approach, it is interesting to see that precisely the non menopause-specific dimensions of the QPI – i.e. the PSY and the PHYS domain – were good predictors of the SF-6D- and EQ-5D-derived utilities. This is not surprising, as the latter are generic instruments and are not able to pick up variance observed in the more condition-specific sexual, menstrual, vasomotor or androgenic dimensions of health of climacteric/postmenopausal women. Research in the area reveals that especially hot flushes – the lead symptom of the menopause – have a significant impact on QoL. The 'wrong' sign of the SEX dimension should not be too worrying at this stage because the significance level is lower compared to the others and probably only reached significance due to the big sample size. In fact, a step-wise addition of variables to the regression model revealed that SEX only slightly improved the model fit. This could be a statistical artefact given that neither the SF-6D nor the EQ-5D

specifically address sexual activity, anyway. The same applies to the VASO (SF-6D and EQ-5D) and the MENS variable (EQ-5D only), although the beta-coefficients do not reach significance in this case.
We intend to undertake a more thorough data analysis such as specification tests (e.g. heterogeneity) and expected-vs.-observed score plots to detect any bias in the model.

This shows that, although especially the SF-6D may be a useful tool to provide preference weights when assessing health in postmenopausal women, a direct valuation survey is needed. Only this will be able to generate overall preference-based utility scores reflecting all menopause-relevant domains of health. Particularly the importance attached to the condition- and treatment-specific QoL domains can thus be elicited. This will also shed light on the extent to which women trade symptom relief and potential side-effects.

The starting point for the latter is a meaningful and robust HSCS. The QPI validation work provides the key items for this. Both factor analysis and the multiple regression confirm the importance of the psychosocial domain, for which a finer scaling is proposed in the HSCS. The vasomotor domain, for the reasons of being the distinctive feature of menopause and highly responsive to treatment, should also be measured on a 5-point (rather than a 3-point) scale in the HSCS. For the other dimensions, the certainly more 'coarse' 3-level scale should suffice in order not to complicate neither the understanding of the health states for respondents (6 dimensions with 5 levels each demands imposing a high cognitive burden) nor the modelling to the analyst. Furthermore, distinctions between levels become fairly artificial beyond 3 levels anyway, as also shown within the face validity testing phase of the QPI.

The next step to generating preference-based single indices involves the valuation survey itself. In terms of the valuation technique, the time trade-off (TTO) has been decided to be somewhat more appropriate. The probabilistic nature of the standard gamble (SG) would seem to artificial in this area where neither the condition itself nor its treatment entail the risk of 'sudden death'. The TTO, in turn, appears to relate more closely to women's perception of life after menopause and therefore requires a lesser amount of imagination; especially in this age group, "time" is probably a more tangible concept than "probability".

REFERENCES

Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 health survey. *J Clin Epidemiol* 1998; 51: 1115-1128

Daly E, Gray A, Barlow D, et al. Measuring the impact of menopausal symptoms on quality of life. *Br Med J* 1993;307:836-40

Dolan P. Modelling valuations for EuroQol health states. *Medical Care* 1997; 35(11): 1095-1108

Gabriel SE, Kneeland TS, Melton LJ, et al. Health-related quality of life in economic evaluations for osteoporosis: whose values should we use? *Med Decis Making* 1999; 19(2): 141-148

Tosteson ANA, Gabriel S, Kneeland T, et al. Has the impact of hormone replacement therapy on health-related quality of life been undervalued? *J Women's Health & Gender-Based Med* 2000;9(2):119-130

Zethraeus N, Johannesson M, Henriksson P, et al. The impact of hormone replacement therapy on quality of life and willingness to pay. *Br J Obstet Gynaecol* 1997; 104: 1191-1195

Zöllner YF, Oliver P, Platts M *et al*. Development of a menopause-specific quality-of-life-questionnaire. Poster and oral presentation.*16th European Congress of Obstetrics and Gynaecology (EAGO/EBCOG)* (abstract n°21, p.14), June 6-9, 2001, Malmö, Sweden.