Article:
Hernández Alava, M., Pudney, S.E. and Wailoo, A.J. (2023) Does EQ-5D Tell the whole story? Statistical methods for comparing the thematic coverage of clinical and generic outcome measures, with application to breast cancer. Value in Health, 26 (9). pp. 1398-1404. ISSN: 1098-3015

https://doi.org/10.1016/j.jval.2023.05.016

ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Methodology

Does EQ-5D Tell the Whole Story? Statistical Methods for Comparing the Thematic Coverage of Clinical and Generic Outcome Measures, With Application to Breast Cancer

Mónica Hernández Alava, PhD, Stephen E. Pudney, MSc, Allan J. Wailoo, PhD

ABSTRACT

Objectives: This study aimed to develop the following: (1) methods for assessing claims in any specific application that a generic outcome measure, such as EQ-5D is deficient in its coverage of 1 or more specified domains, and (2) a simple method of judging whether any such deficiency is likely to be quantitatively important enough to call into question evaluations based on the generic instrument. Also to demonstrate the applicability of the methods in the important area of breast cancer.

Methods: The methodology requires a data set with observations from a generic instrument (eg, EQ-5D) and also a more comprehensive clinical instrument (eg, FACT-B [Functional Assessment of Cancer Therapy – Breast]). A standardized 3-component statistical analysis is proposed for investigating the claim that the generic measure inadequately captures some specified dimension covered by the latter instrument. A theoretically based upper bound on the bias induced by deficient coverage is derived based on the assumption that the designers of the (*k*-dimensional) generic instrument did succeed in identifying the *k* most important domains.

Results: Data from the MARIANNE breast cancer trial were analyzed and results suggested that impacts on personal appearance and relationships may be inadequately represented by EQ-5D. Nevertheless, the indications are that the bias in quality-adjusted life-year differences from deficient coverage by EQ-5D is likely to be modest.

Conclusions: The methodology offers a systematic approach to determining whether there is clear evidence consistent with any claim that a generic outcome measure such as EQ-5D misses an important specific domain. The approach is readily implementable using data sets that are available in many randomized controlled trials.

Keywords: breast cancer, cost-effectiveness, EQ-5D.

VALUE HEALTH. 2023; 26(9):1398–1404

Introduction

For economic evaluation of medical technologies, consistency is supremely important— consistency across disease areas, patient groups, and health technologies. That need has led to the development of generic (rather than disease-specific) instruments for measuring health outcomes, with corresponding value sets for the calculation of quality-adjusted life-years (QALYs). The consistency achieved by generic instruments comes at some cost because they may not adequately cover all dimensions of health benefits relevant in all situations.

Organizations such as the National Institute for Health and Care Excellence explicitly recognize this. National Institute for Health and Care Excellence expresses a preference for EQ-5D, with decisions to deviate requiring supporting empirical evidence.[1] In practice, claims that a particular instrument such as EQ-5D is inappropriate are often based on qualitative arguments that certain symptoms of the disease in question do not feature in its classification system. Most objections to the coverage of generic instruments do not come with a theoretically based evaluation of the bias in economic evaluation caused by these limitations.

The adequacy of a generic instrument cannot be judged solely through thematic questionnaire analysis because, even if it omits items dealing with a specific aspect of health, it may still capture that aspect indirectly. For example, if the impact of disease and treatment impairs the quality of personal relationships (not covered by EQ-5D), that may, in turn, have indirect impacts on usual activities and on anxiety/depression. But EQ-5D covers both; therefore, the relationships aspect may be captured indirectly.

We propose methods for investigating the conceptual completeness of generic instruments such as EQ-5D, with equal applicability to any other generic instrument with utility scoring systems. Such generic measures may be conceptually incomplete in one context (in which there is a substantial and differential impact of treatment in excluded health outcome dimensions), but

not in others (in which the effects of treatment are uniform in those dimensions).

Our aims are, first to develop simple econometric methods for application to patient-level data to test the validity of claims that a given generic outcome measure underrepresents aspects of the patient impact of disease.

Second, we detail the circumstances in which a cost-effectiveness analysis of alternative health technologies would be substantially distorted by the generic instrument's incomplete coverage and develop a simple revaluation argument to indicate the potential scope for bias.

Third, we apply the methodology to data from a trial in the area of female patients with breast cancer, in which there is a priori reason for concern about the coverage of the EQ-5D instrument, using the disease-specific FACT-B (Functional Assessment of Cancer Therapy – Breast) instrument to identify potential omissions.

Methods

Statistical Methods

We start from 2 basic principles:

- P1 There should be face validity for the claimed specific omission from the generic instrument. Purely empirical "data dredging" approaches can lead to spurious results arising by chance alone and should be avoided.
- P2 The claim should be supported by evidence from comprehensive data that cover both the content of the generic instrument and any health dimensions claimed to be missing from it.

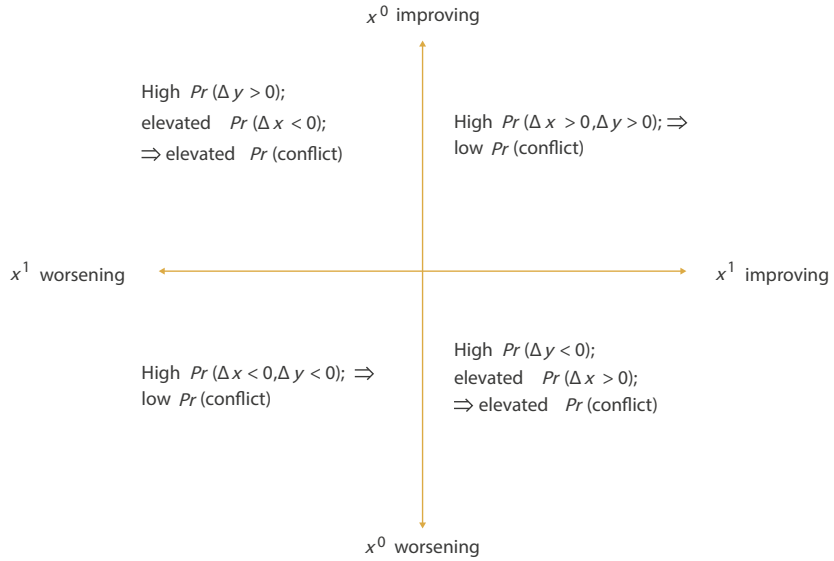
Consider a data set covering individuals $i = 1...\ n$, observed repeatedly in waves of measurement $t = 1, 2,...$ Two instruments measure the health-related quality of life outcome. The clinical instrument covers a suitably wide range of dimensions and yields a set of patient outcome measures denoted $X_{1it},\ldots,X_{Jit}$. The generic instrument generates fewer data items $Y_{1it}\ldots Y_{5it}$. For the generic instrument, a utility score y_{it} is calculated at each measurement, as a function of the Y_{jit}. No such utility score exists for the clinical instrument.

Principle P2 allows us to divide the items of the clinical instrument into 2: a set of *noncore* items $X^1_{1it}\ldots X^1_{J^1it}$, measuring aspects claimed to be omitted, and a set of *core* items $X^0_{1it}\ldots X^0_{J^0it}$ presumed to be covered directly by the generic instrument. In most cases (as we do here), all items in the clinical instrument, except for those classed as noncore, will be included in the core, although there may sometimes be a case for excluding some items completely. We construct an overall clinical index x_{it}, and separate subindexes, x^0_{it} and x^1_{it}, for the core and noncore items. Common practice is to use sums of the relevant items possibly after suitable rescaling, but alternative factor methods may be used.

If the generic measure is conceptually incomplete, we should find evidence consistent with the following hypotheses:

- H1 Responses to core questions in the clinical instrument are more highly correlated with the generic utility score or its constituent items than are responses to noncore questions.
- H2 Hypothesis H2 extends H1 by requiring that, in a joint multivariate modeling context, the core clinical items $X^0_{1it}\ldots X^0_{J^0it}$ should have much greater joint predictive power for the generic instrument y_{it} than the noncore items $X^1_{1it}\ldots X^1_{J^1it}$.
- H3 The probability of conflict between the direction of change indicated by the overall clinical index $(x_{it-1} \rightarrow x_{it})$ and the generic utility $(y_{it-1} \rightarrow y_{it})$ will be greater for individuals with core and noncore dimensions of health changing in different directions. Figure 1 sets out the detail of H3: if the core measure x^0 improves between 2 waves of measurement, we expect the generic measure y to be improving also; if, in addition, the noncore dimension x^1 is improving (ie, the NE quadrant), the overall clinical index x must be improving, and there is only a small probability of directional conflict between the measures x and y. But if the noncore dimension x^1 is worsening (ie, the NW quadrant), and does so enough to alter the direction of change in the overall index x, there will be an elevated probability of conflict because the generic index y tends to track the core x^0 rather than the full clinical index. The converse argument holds for the SW and SE quadrants of Figure 1. Note that a conflict

Figure 1. Probabilities of conflicting directions of change in generic and clinical measures by direction of change in core and noncore dimensions. The symbol Δ denotes the change in level between two waves of measurement.



between the direction of change in the core and noncore components of FACT-B does not necessarily imply that EQ-5D is deficient in the noncore domain—instruments such as FACT-B and EQ-5D may capture effects indirectly, and the different designs of questions in the FACT-B core and EQ-5D mean that they may differ in the way they do so.

Our proposal is conservative: given the large existing body of consistent decision making using generic instruments, an accepted instrument such as EQ-5D should not be called into question lightly. In any specific context there should be a clear a priori conceptual case, supported by statistical evidence consistent with H1 to H3. There is a wide choice of statistical techniques to turn these hypotheses into operational procedures, but methods should be simple and capable of routine use in public decision making. We make specific recommendations below, using the important example of breast cancer to illustrate the approach.

A Bound on the QALY Bias

If a health dimension omitted from the generic instrument is seen as important by the general public, there is potential for bias in public decision making. A cost-effectiveness analysis contrasting 2 treatments $T = 0$ and $T = 1$ is built up from the net present value of utility differences, $E(y|T=1) - E(y|T=0)$ over the treatment period. Bias occurs when this measured difference does not coincide with $E(v|T = 1) - E(v|T = 0)$, in which v is the "true" utility score. There are 2 key issues: anchoring bias and differential treatment bias (the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016 contains a formal analysis of the bias).

Anchoring bias concerns the notion of perfect health, conventionally fixed at utility value 1. Responses to the generic instrument are converted into values by conducting valuation experiments (eg, using time trade-off). Such methods present experimental participants with alternative hypothetical sets of information on the items covered by the instrument and ask them to make choices between alternatives. If important domains are missing, participants necessarily make implicit assumptions about the levels of the missing dimensions, but we do not know what those implicit assumptions are. The best observable state will wrongly define perfect health if any omitted dimension is assumed by participants to be at a suboptimal level. Nevertheless, anchoring bias seems unlikely to be very important in practice. If participants in valuation experiments assume perfect health with respect to noncore dimensions, anchoring bias is 0. Moreover, if it does exist, its effect will be uniform across treatments; therefore, it will not (on its own) alter the QALY rankings of alternative treatments.

Differential treatment bias arises when the prediction $E(y|T)$ departs from $E(v|T)$ in different ways for each treatment T. It is more complex than anchoring bias, and its effect depends on how the treatments being compared interact with the omitted dimensions. For example, our empirical breast cancer application suggests that impacts on appearance and relationships are potentially understated by EQ-5D. If we compare treatments involving the same adverse effects on appearance/relationships, the limitations of EQ-5D will not distort the analysis. But if we use EQ-5D to compare, say, radical mastectomy with local excision, or forms of chemotherapy causing substantially different degrees of hair loss, the results will be biased in favor of treatments that are seen by patients as more disfiguring.

The obvious response to findings of incomplete coverage is to develop an expanded measure with more dimensions and carry out new valuation experiments to produce an associated value set—a costly and time-consuming option. "Bolt-ons"[2,3] are a half-way house that involve special-purpose extensions, rather than wholesale replacement of the generic instrument, but still require new valuation exercises.

We propose an intermediate step, using estimates interpretable as an upper bound on the magnitude of the bias in measured QALYs to indicate whether such extensions are likely to make a practical difference. We start from the reasonable assumption that the original designers of the generic instrument were thorough in their background research and correctly identified the dimensions seen by the public as being most important. If that is so, any dimension excluded from the instrument can be no more important to the public than the least important included dimension. The details of this depend on the particular generic instrument and value set, and we develop it further in the EQ-5D/breast cancer application.

Data

Cancer is an area in which claims of limitations of EQ-5D have been made.[4] The FACT-B instrument provides a patient-reported outcome measure for treatment of breast cancer in women.[5] It comprises 37 items measured on a 0 to 4 response scale, which are summed to give the overall summary measure x_{it}, after normalisation so that higher values indicate better outcomes. FACT-B covers more dimensions than generic measures, notably 3 not directly covered by EQ-5D: Appearance (items B2-B5, B8, B9), Relationships (items GS1-GS7), and Sleep (item GF5). In the spirit of placebo testing, we also separate out Work (items GF1 and GF2), which is of great policy interest but could be expected to be covered adequately by the usual activities dimension of EQ-5D (see Appendix Table A1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016 for further detail).

Our data come from the phase III MARIANNE clinical trial,[6,7] which compared 3 treatments for HER2-positive advanced breast cancer among 1095 patients interviewed repeatedly over a 3-year period. The trial was approved by relevant institutional review boards or independent ethics committees at each site. Patients provided written informed consent. Both FACT-B and the original EQ-5D-3L were measured at each interview. After excluding cases with missing responses, our analysis sample comprises 888 individuals, with an average of 12 measurements each. For the core index of FACT-B, x^0, we sum all 21 items not used to construct the indexes for Appearance, Relationships, Sleep, and Work. The FACT-B index x_{it} is constructed in the standard way,[5] and the Dolan value set[8] generates the EQ-5D utility score y. Note that there are established subindexes of FACT-B,[5] but our division into core and noncore indexes (x^0_{it} and x^1_{it}) has a different purpose—it captures claimed conceptual distinctness from EQ-5D.

Results

Hypothesis H1

Use of indexes in a correlation analysis to investigate the hypothesis H1 risks a potential pitfall. A subindex constructed from a large number of items will tend to average out more of the randomness in responses than another subindex summing fewer items. Thus, a core index x^0_{it} covering more items than a noncore index x^1_{it} may spuriously conform to hypothesis H1 by virtue of its greater number of components and reduced measurement noise. Therefore, we examine the average correlation of the items making up the index, rather than the correlation of indexes, and also correlations of wave-to-wave changes because cost-effectiveness

analysis is concerned with improvements in health states, necessarily involving change over time. Moreover, different individuals may interpret response scales differently, which could distort comparisons between individuals, whereas within-individual changes over time remain meaningful.

The last column of Table 1 shows that core FACT-B items are on average significantly (at the 95% confidence level) more strongly correlated with the EQ-5D utility score than are the noncore groups of items for Appearance, Relationships, and Sleep. This is true for both levels and changes. The preceding 5 columns in Table 1 show correlations between each of the 5 items of the EQ-5D health description and the core and noncore items of FACT-B. Items in the Appearance and Relationships categories of FACT-B are less correlated with all the EQ-5D items than are the core items of FACT-B—significantly so (at the 1% level) in all of the 5 EQ-5D domains. Thus, hypothesis H1 holds empirically for both. There is no significant difference in the anxiety/depression domain for the single-item Sleep aspect; therefore, EQ-5D goes some way toward capturing impacts on sleep.

For the Work dimension, there are average correlations of 0.457 (levels) and 0.104 (changes) between the EQ-5D utility score and the Work items. The former actually exceeds the average core correlation (0.418), whereas the latter is less than its comparator 0.132, but insignificantly so at the 95% level. There is some debate about the role of productivity losses in cost-effectiveness analysis and the risk of double counting those losses. There are 2 separate issues: do the dimensions covered by generic instruments adequately cover impacts on work activity with productivity-limiting effect? Do health-related quality of life valuations provided by choice experiments adequately capture the value of those impacts through implicit assumptions by participants about earnings loss? Most of the relevant research literature[9–11] relates to the latter question, which is not our concern here. The results in Table 1 are important for the former, and suggest that the EQ-5D health description adequately captures impacts on work, primarily through the usual activities domain.

Hypothesis H2

Hypothesis H2 predicts that, in a joint analysis using information from all the clinical items to model the EQ-5D utility score, the core items X^0_{jit} will give much more predictive power than the noncore items X^1_{jit}. A natural statistical framework is a latent variable (LV) structure, in which the core and noncore dimensions are represented by unobservable LVs, with the observable questionnaire items acting as "noisy" indicators of those latent dimensions. This LV approach has 4 advantages over the usual method of constructing indexes by summing items: it allows for the ordinal nature of FACT-B items, it accommodates random measurement "noise" in the item responses, it allows for nonuniform sensitivity by estimating different sensitivity parameters (factor loadings) for each item, and, unlike a 2-stage approach that first constructs core and noncore indexes then uses them to model EQ-5D, it integrates the construction of indexes and modeling of EQ-5D into a single, more efficient, procedure. Estimation is by maximum likelihood and the estimates can be used to construct Empirical Bayes indexes as the expectations of the LVs conditional on the observed questionnaire items. Such indexes use all available information and give appropriately higher weight to items that are relatively more informative. Technical details are in the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016.

The evidence on hypothesis H1 is strongest for the Appearance and Relationships domains; therefore, we estimate an LV model

Table 1. Average correlations of items of the core and noncore FACT-B with the EQ-5D utility score and its components, pooled sample.

FACT-B subindex	EQ-5D dimension					EQ-5D utility score
	Mobility	Self-care	Usual activities	Pain	Anxiety / depression	
Correlations of responses (X^0_{jit}, X^1_{jit}) with EQ-5D						
Core FACT-B	−0.259	−0.220	−0.316	−0.359	−0.330	0.418
Appearance [*P*-value]	−0.127 [.000]	−0.141 [.000]	−0.163 [.000]	−0.165 [.000]	−0.174 [.000]	0.210 [.000]
Relationships [*P*-value]	−0.060 [.000]	−0.072 [.000]	−0.083 [.000]	−0.095 [.000]	−0.161 [.000]	0.133 [.000]
Work [*P*-value]	−0.361 [.000]	−0.298 [.000]	−0.497 [.000]	−0.323 [.068]	−0.317 [.519]	0.457 [.056]
Sleep [*P*-value]	−0.182 [.000]	−0.146 [.002]	−0.236 [.000]	−0.303 [.002]	−0.348 [.322]	0.350 [.001]
Correlations of response changes ($\Delta X^0_{jit}, \Delta X^1_{jit}$) with changes in EQ-5D						
Core FACT-B	−0.068	−0.043	−0.079	−0.103	−0.081	0.132
Appearance [*P*-value]	−0.019 [.000]	−0.031 [.141]	−0.034 [.000]	−0.035 [.000]	−0.036 [.000]	0.057 [.000]
Relationships [*P*-value]	−0.009 [.000]	−0.014 [.022]	0.001 [.000]	0.008 [.000]	−0.030 [.000]	0.011 [.000]
Work [*P*-value]	−0.079 [.382]	−0.050 [.560]	−0.115 [.005]	−0.051 [.000]	−0.046 [.002]	0.104 [.032]
Sleep [*P*-value]	−0.034 [.021]	−0.015 [.040]	−0.059 [.157]	−0.078 [.036]	−0.072 [.494]	0.097 [.021]

Note. P-values in square brackets are for tests of the hypothesis of a zero difference between the average core correlation (top row of table) and average noncore correlation; 2-sided confidence intervals from bootstrap simulations (500 replications).
FACT-B indicates Functional Assessment of Cancer Therapy – Breast.

Table 2. Estimated impacts of core and noncore FACT-B indexes on EQ-5D utility values: two variants of the LV model and fixed-effects regression and random effects Tobit models of summation indexes.

FACT-B component	LV model*		LV model for between-wave response changes*	Models for summation indexes†	
	Non-Tobit	Tobit		FE regression	RE Tobit
Core	0.179‡ (0.008)	0.250‡ (0.013)	0.058‡ (0.006)	0.111‡ (0.003)	0.121‡ (0.003)
Appearance	−0.014§ (0.007)	−0.026‡ (0.009)	0.006 (0.008)	0.014‡ (0.003)	0.012‡ (0.003)
Relationships	−0.032‡ (0.006)	−0.043‡ (0.008)	−0.005 (0.003)	−0.012‡ (0.003)	−0.015‡ (0.003)
Work	-	-	-	0.015‡ (0.002)	0.023‡ (0.003)
Sleep	-	-	-	0.003 (0.002)	0.007‡ (0.002)

FACT-B indicates Functional Assessment of Cancer Therapy – Breast; FE, fixed effects; LV, latent variable; RE, random effects.
Statistical significance:
‖10%.
*3-factor model: coefficients of impact of LVs for FACT-B core, appearance, and relationships on EQ-5D utility.
†Marginal responses for regression models; indexes are in SD units; Tobit variants allow for an upper limit at EQ-5D = 1; SEs in parentheses.
‡1%.
§5%.

incorporating those 2 dimensions alongside the core. The first 3 columns of Table 2 present estimates of the key impacts of the LVs for core FACT-B, and noncore Appearance and Relationships (full parameter estimates in Appendix Table A2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016). They show the marginal effect on the EQ5D score of a 1-SD increase, using versions of the model in levels, without and with Tobit allowance for the upper limit on the EQ-5D utility score, and a version in time-difference form. For comparison, the last 2 columns of Table 2 show models based on summation indexes of the core items of FACT-B and noncore items covering Appearance, Relationships, Work, and Sleep. These last 2 models are respectively fixed-effects linear regression (allowing for persistent individual differences in response styles) and random effects Tobit (additionally allowing for the limit of 1.0 on EQ-5D utility scores).

In each case, the association of EQ-5D with noncore components of FACT-B are weak (always less than one-fifth the magnitude) compared with the association with core FACT-B. For the time-difference model (arguably the most reliable because it eliminates persistent individual differences in response behavior) the association of EQ-5D with the core is highly statistically significant, whereas the latent Appearance and Relationships components show no significant association.

Hypothesis H3

Hypothesis H3 predicts a higher probability of conflict between FACT-B and EQ-5D in terms of their directions of change over time when the core outcome is improving and the noncore outcome worsening over time, or vice versa (NW and SE quadrants of Figure 1). Using summation indexes, Appendix Table A3 of in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016 shows that this is the case empirically, for each of the Appearance, Relationships, Work, and Sleep domains, in turn, with differences larger for Appearance and Relationships than for Work and Sleep. Table 3 shows results of a more comprehensive joint analysis using probit models (because conflict/nonconflict is a binary distinction), with random effects allowing for persistent

Table 3. Random effects probit marginal effects of the direction of change in noncore aspects on the probability of conflicting directions of change in FACT-B and EQ-5D.

Non-core domain	LV Tobit indexes		LV difference indexes		Summation indexes	
	Core improving*	Core worsening†	Core improving*	Core worsening†	Core improving‡	Core worsening§
Appearance	0.089£ (0.007)	0.073£ (0.007)	0.071£ (0.007)	0.062£ (0.006)	0.044£ (0.005)	0.051£ (0.006)
Relationships	0.084£ (0.007)	0.078£ (0.007)	0.065£ (0.006)	0.071£ (0.008)	0.060£ (0.006)	0.052£ (0.006)
Work	-	-	-	-	0.027£ (0.005)	0.028£ (0.005)
Sleep	-	-	-	-	0.026£ (0.006)	0.018£ (0.006)

Note. Average marginal effects of conflicting direction of change in core and noncore indexes on the probability of conflict between EQ-5D and FACT-B. Standard errors in parentheses.
FACT-B indicates Functional Assessment of Cancer Therapy – Breast; LV, latent variable.
Statistical significance:
‖10%.
¶5%.
£1%.
*From estimates of model (A5) in the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016.
†From estimates of model (A6) in the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016.
‡From estimates of extension of model (A5) in the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016 to include covariates for Work and Sleep.
§From estimates of extension of model (A6) in the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016 to include covariates for Work and Sleep.

differences in individual reporting styles. They are shown as average marginal effects of conflict in the direction of change in core and noncore indexes on the probability of conflict between the EQ-5D score and overall FACT-B (see Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016 for details). Three alternative versions of the indexes x_{it}^0 and x_{it}^1 are used: Empirical Bayes indexes from the Tobit LV model in levels and the LV model in differences, and simple summation indexes (the LV model provided indexes for Appearance and Relationships only). All estimates indicate that dissonant changes in the Relationships and Appearance aspects of FACT-B are strongly and significantly associated with conflicting directions of change in overall FACT-B and EQ-5D. The effects are large: if the Relationships or Appearance index is moving in the opposite direction to the core, it increases the average conflict rate by 4-9 percentage points (relative to a base level of 8.6%), depending on the model used. Using summation indexes, the effects of Work and Sleep are smaller, ranging from 1.8 to 2.8 percentage points.

A Bound on the Bias in Mean Utility Differences

How sensitive is measurement of utility differences to the omission of the Appearance and Relationships domains from EQ-5D? The structure of the EQ-5D utility tariff for the United Kingdom[9] implies unambiguously that the "usual activities" domain is the least important of the 5. Changing "no problems with performing my usual activities" to "some problems..." reduces the score by 0.036, whereas a change to "unable to..." reduces it by 0.363 (or 0.094 if another domain is reported at the worst level 3). To extend the formula, we put the noncore clinical index x^1 in the same 3-level format as the EQ-5D items and incorporate it into EQ-5D to give an upper bound, under 3 assumptions:

- A1 The extended valuation formula has the same structure (linear with extreme state adjustments) as the standard tariff.
- A2 The utility decrements associated with worsening in the new domain are identical to those associated with the usual activities domain of EQ-5D.
- A3 Participants in the TTO valuation experiments underlying the Dolan formula implicitly assumed no difficulties in the new domain.

We also need a procedure for converting the Appearance and Relationships indexes into the 3-level format required by EQ-5D-3L. The resulting formula is then used to construct the modified QALY difference (see Figures A1 and A2 and technical details, in the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016). Without direct evidence on the distribution of responses that an EQ-5D style 3-choice question would produce, we use the distributions of responses to the relevant FACT-B items to define cut-offs used to convert the continuous Appearance and Relationships indexes into 3-level form. As a check on sensitivity, this is done in 2 alternative ways, under a "broad" interpretation treating FACT-B categories 1-3 as central and a "narrow" interpretation treating only category 2 as central. Table 4 and the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.05.016 give details.

The MARIANNE data set available to us has no trial arm identifiers; therefore, we cannot revisit the MARIANNE economic evaluation. Nevertheless, there are large differences in measured outcomes between groups defined by baseline cancer stage; therefore, we can trial the proposed approach by examining how the comparison across cancer stage groups of mean EQ-5D utility scores is affected when we adjust for possible effects on appearance and relationships. If the measured differences are not robust, then EQ-5D should be regarded as potentially suspect in any trial of interventions with differing effects on appearance and relationships.

Table 4 compares standard and adjusted EQ-5D outcomes for patients in the MARIANNE trial with stage I and stage II cancer at baseline. Mean directly measured EQ-5D for stage I patients is 0.813; for stage II patients outcomes are worse by 0.075 on average. The effect of extending EQ-5D valuations for appearance

Table 4. Comparison of mean EQ-5D scores by cancer stage at baseline, with standard EQ-5D and extensions for appearance or relationships.

	Measured EQ-5D	EQ-5D extended for			
		Appearance		Relationships	
		Simple index*	LV index†	Simple index*	LV index†
Discretization with broad central category (extreme category proportions 0.037, 0.46)‡					
Mean level of EQ-5D in stage I group	0.813 (0.021)	0.793 (0.022)	0.793 (0.022)	0.797 (0.022)	0.797 (0.023)
Stage II vs stage I	−0.075§ (0.026)	−0.081§ (0.027)	−0.083§ (0.027)	−0.081§ (0.027)	−0.079§ (0.027)
Discretization with narrow central category (extreme category proportions 0.10, 0.74)‖					
Mean level of EQ-5D in stage I group	0.813 (0.021)	0.794 (0.023)	0.792 (0.023)	0.803 (0.022)	0.803 (0.022)
Stage II vs stage I	−0.075§ (0.026)	−0.084§ (0.027)	−0.084§ (0.028)	−0.083§ (0.027)	−0.082§ (0.027)

Note. Bootstrap standard errors in parentheses (clustered by individual; 500 replications).
FACT-B indicates Functional Assessment of Cancer Therapy – Breast; LV, latent variable.
Statistical significance of stage I vs stage II difference:
£5%.
*Noncore index constructed as unweighted sum of items.
†Noncore index constructed as Empirical Bayes estimate from LV model.
‡Sample proportions of extreme categories fixed at mean proportions of relevant FACT-B responses of 0 and 4.
§1%.
‖Sample proportions of extreme categories fixed at mean proportions of relevant FACT-B responses of 0 or 1 and 3 or 4.
Statistical significance of stage I vs stage II difference:
¶10%.

or relationships is to reduce the mean measured score by an amount ranging from 0.010 to 0.021, depending on the index and discretization.

Nevertheless, for cost-effectiveness work, differences rather than levels matter, and the extension to EQ-5D increases the magnitude of differences between stage I and II patients by only 0.004-0.009. These are modest measurement changes—less than one-eighth the width of a 95% confidence interval for the difference in standard EQ-5D, and less than one-twentieth of the SD of the stage I standard EQ-5D score. In this application, the omission of the Appearance and Relationships domains from EQ-5D seems likely to be a minor source of bias in cost-effectiveness work.

Summary and Conclusions

Generic preference-based measures are widely used to value health states and estimate QALYs. It is sometimes argued that those instruments are inappropriate for measuring health benefit in some situations, but there remains uncertainty about which disease areas, types of impairment or symptoms are poorly captured and how frequently such issues arise. Many claims have been made but their validity is difficult to assess.[4] This study proposes easily implementable methods designed to assist in assessing claims that specific aspects of health are omitted or undervalued, using patient-level responses.

In a case study assessing EQ-5D-3L relative to the FACT-B outcome measure for women with breast cancer, we find that the impact of ill health on Appearance and Relationships is not well captured by EQ-5D. Technologies that differ from comparators in these areas may be significantly misvalued in cost-effectiveness studies. The impact on Sleep, also not directly measured by EQ-5D, may be a less serious omission, whereas—as expected – there is little evidence of any problem relating to Work.

This is a first application demonstrating the potential value of new methods, and further applications are needed in other health areas and other outcome measures than EQ-5D. It is also important to ensure that the use of these methods is not restricted to the assessment of new technologies alone, without also acknowledging that the issue of imperfect benefit assessment applies to existing services that could be displaced. Nevertheless, the developments here demonstrate that decision makers faced with potentially incomplete measures of outcome can be provided with information to judge those claims and simple indicators of the potential for bias in decision making. In cases which incompleteness turns out to entail negligible bias, this approach avoids the need to embark on costly, long-term alternative benefit valuation. In other cases, it may provide a means of establishing that there are strong arguments for some form of "bolt-on" study.[2,12,13]

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2023.05.016.

Article and Author Information

Accepted for Publication: May 24, 2023

Published Online: June 26, 2023

doi: https://doi.org/10.1016/j.jval.2023.05.016

Author Affiliations: School of Health and Related Research, University of Sheffield, Sheffield, England, UK (Alava, Pudney, Wailoo).

Correspondence: Stephen E. Pudney, MSc, School of Health and Related Research, University of Sheffield, 30 Regent St, Sheffield S1 4DA, England, United Kingdom. Email: steve.pudney@sheffield.ac.uk

Author Contributions: *Concept and design*: Alava, Pudney, Wailoo
Acquisition of data: Alava, Wailoo
Analysis and interpretation of data: Alava, Pudney, Wailoo
Drafting of the article: Alava, Pudney, Wailoo
Statistical analysis: Alava, Pudney, Wailoo
Obtaining funding: Alava, Wailoo

Conflict of Interest Disclosures: Drs Alava, Pudney, and Wailoo reported receiving grants from the National Institute for Health and Care Excellence during the conduct of the study. Dr Wailoo is an editor for *Value in Health* and had no role in the peer-review process of this article. No other disclosures were reported.

Funding/Support: This article is based on an unpublished report which was funded by the National Institute for Health and Care Excellence. F. Hoffman La Roche made available MARIANNE trial data.

Role of Funder/Sponsor: The National Institute for Health and Care Excellence provided funding for the stud. F. Hoffman La Roche gave access to the data but had no role in the design and conduct of the study; management, analysis, and interpretation of the data; preparation, review or approval of the article; and decision to submit the article for publication. Staff members at National Institute for Health and Care Excellence provided comments on an initial draft of the article.

Acknowledgment: Rosie Lovett and Alan Lamb of National Institute for Health and Care Excellence made helpful comments, but the views expressed here, and any errors or omissions, are those of the authors only. National Institute for Health and Care Excellence may take account of part or all of the research described here if it considers it appropriate, but it is not bound to do so.

REFERENCES

1. Guide to the methods of technology appraisal 2013. National Institute for Health and Care Excellence. https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781. Accessed February 6, 2020.
2. Yang Y, Rowen D, Brazier J, Tsuchiya A, Young T, Longworth L. An exploratory study to test the impact on three "bolt-on" items to the EQ-5D. *Value Health*. 2015;18(901):916.
3. Geraerds AJLM, Bonsel GJ, Janssen MF, Finch AP, Polinder S, Haagsma JA. Methods used to identify, test, and assess impact on preferences of bolt-ons: a systematic review. *Value Health*. 2021;24(6):52–60.
4. Wailoo AJ, Davis S, Tosh J. The incorporation of health benefits in cost utility analysis using the EQ-5D: report by the decision support unit 2010. https://www.sheffield.ac.uk/nice-dsu/methods-development/eq-5d. Accessed September 14, 2022.
5. Brady M, Cella D, Mo F, et al. Reliability and validity of the Functional Assessment of Cancer Therapy-Breast quality-of-life instrument. *J Clin Oncol*. 1997;15(3):974–986.
6. Perez EA, Barrios C, Eiermann W, et al. Trastuzumab emtansine with or without pertuzumab versus trastuzumab plus taxane for human epidermal growth factor receptor 2 positive, advanced breast cancer: primary results from the phase III MARIANNE Study. *J Clin Oncol*. 2017;35(2):141–148.
7. Perez EA, Barrios C, Eiermann W, et al. Trastuzumab emtansine with or without pertuzumab versus trastuzumab with taxane for human epidermal growth factor receptor 2–positive advanced breast cancer: final results from MARIANNE. *Cancer*. 2019;125(22):3974–3984.
8. Dolan P. Modeling valuations for EuroQol health states. *Med Care*. 1997;35(11):1095–1108.
9. Tilling C, Krol M, Tsuchiya A, Brazier J, Brouwer W. In or out? Income losses in health state valuations: a review. *Value Health*. 2010;13(2):298–305.
10. Tilling C, Krol M, Tsuchiya A, Brazier J, van Exel J, Brouwer W. Does the EQ-5D reflect lost earnings? *Pharmacoeconomics*. 2012;30(1):47–61.
11. Shiroiwa T, Fukuda T, Ikeda S, Shimozuma K. QALY and productivity loss: empirical evidence for "double counting,". *Value Health*. 2013;16(4):581–587.
12. Krabbe PFM, Stouthard MEA, Essink-Bot ML, Bonsel GJ. The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system. *J Clin Epidemiol*. 1999;52(4):293–301.
13. Yang Y, Brazier J, Tsuchiya A. Effect of adding a sleep dimension to the EQ-5D descriptive system: a "bolt-on" experiment. *Med Decis Making*. 2014;34(1):42–53.