

Original citation:

Petrou, Stavros, Kwon, J. and Madan, Jason (2018) A practical guide to conducting a systematic review and meta-analysis of health state utility values. *PharmacoEconomics*, 36 (9). pp. 1043-1061. doi:10.1007/s40273-018-0670-1

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/101694>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

"The final publication is available at Springer via <https://doi.org/10.1007/s40273-018-0670-1>"

A note on versions:

The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP url' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

A practical guide to conducting a systematic review and meta-analysis of health state utility values

Short running head: Systematic review and meta-analysis of health utilities

Authors: Stavros Petrou,^a Joseph Kwon,^b Jason Madan^a

^a Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry
CV4 7AL, UK.

^b School of Health and Related Research, The University of Sheffield, 30 Regent St,
Sheffield S1 4DA, UK

Contact for correspondence:

Professor Stavros Petrou

Warwick Medical School

University of Warwick

Coventry

CV4 7AL

UK

Tel: 0044 (0)2476 151124

FAX: 0044 (0)2476 151586

E-mail: S.Petrou@warwick.ac.uk

Abstract

Economic analysts are increasingly likely to rely on systematic reviews and meta-analyses of health state utility values to inform the parameter inputs of decision-analytic modelling based economic evaluations. Beyond the context of economic evaluation, evidence from systematic reviews and meta-analyses of health state utility values can be used to inform broader health policy decisions. This paper provides practical guidance on how to conduct a systematic review and meta-analysis of health state utility values. The paper outlines a number of stages in conducting a systematic review, including identifying the appropriate evidence, study selection, data extraction and presentation, and quality and relevance assessment. The paper outlines three broad approaches that can be used to synthesise multiple estimates of health utilities for a given health state or condition, namely fixed-effect meta-analysis, random-effects meta-analysis, and mixed-effects meta-regression. Each approach is illustrated by a synthesis of utility values for a hypothetical decision problem, and software code is provided. The paper highlights a number of methodological issues pertinent to the conduct of meta-analysis or meta-regression. These include the importance of limiting synthesis to ‘comparable’ utility estimates, e.g. those derived using common utility measurement approaches and sources of valuation; reliance on limited or poorly reported published data in primary utility assessment studies; the use of aggregate outcomes within analyses; approaches to generating measures of uncertainty; handling of median utility values; challenges surrounding the disentanglement of utility estimates collected serially within the context of prospective observational studies or prospective randomised trials; challenges surrounding the disentanglement of intervention effects; and approaches to measuring model validity. Areas of methodological debate and avenues for future research are highlighted.

Key Points for Decision Makers

- The process of systematically reviewing health state utility values should involve a number of formal stages, including identifying the appropriate evidence, study selection, data extraction and presentation, and quality and relevance assessment.
- When there are multiple estimates of health utilities for a given health state, fixed-effect meta-analysis, random-effects meta-analysis and mixed-effects meta-regression are alternative approaches for pooling values collected across a number of studies.
- There are a number of methodological issues pertinent to the conduct of meta-analysis or meta-regression, including for example the importance of limiting synthesis to utility estimates derived using comparable methods and the challenges raised by limited or poorly reported data in primary utility assessment studies.

1. Introduction

Cost-utility analysis remains the preferred form of economic evaluation for health technology assessment (HTA), pricing and reimbursement authorities in several countries, including the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia [1], the Canadian Agency for Drugs and Technologies in Health (CADTH) in Canada [2], the Haute Autorité de Santé (HAS) in France [3], the College voor zorgverzekeringen (CVZ) in the Netherlands [4], the CatSalut in Spain (Catalonia) [5], the National Institute of Health and Care Excellence (NICE) in England and Wales [6], and the Scottish Medicines Consortium (SMC) in Scotland [7]. The results of cost-utility analyses are commonly expressed in terms of incremental cost per quality-adjusted life year (QALY) gained. In order to generate QALY estimates, preference-based health-related quality of life weights, commonly referred to as health state utility values, are combined with data on length of time in the health states of interest. Notably, utility values reflect people's preferences or social judgements about the relative worth of alternative health states. They therefore move beyond a narrow biomedical perspective on health outcomes measurement towards an extra-welfarist perspective that can inform allocative decision-making.

Health economists apply a number of approaches for estimating health state utility values. These include direct valuation methods, such as the standard gamble (SG), time trade-off (TTO) and visual analogue scale (VAS); multi-attribute health status classification systems with preference scores, such as the EQ-5D [8], Health Utilities Index (HUI) [9], or SF-6D [10]; mapping from non-preference-based measures onto generic preference-based measures of health; development of preference-based measures derived from existing non-preference-based measures; and development of new preference-based measures encompassing *de novo* descriptive systems and utility algorithms [11]. In a single study economic evaluation, for example a within-trial economic evaluation, alternative approaches for estimating health state

utility values can be prospectively incorporated into the study design. However, single study economic evaluations often present methodological challenges to the optimal collection of health state utility values as a result of, for example, the timing and frequency of health utility assessments and the heterogeneity of the study sample [12]. Moreover, it remains relatively rare that a single study economic evaluation will generate all the data required to inform a policy decision [13]. In the context of decision-analytic modelling based economic evaluations, analysts generally lack the time and resources to estimate primary utility values for all health states of interest. There are several circumstances therefore where analysts will resort to reviews of external evidence on health state utility values.

A number of structured or systematic reviews of health state utility values have been reported in the literature, the results of which have acted as data inputs into economic evaluations. Early seminal research by Tengs and Wallace identified 1,000 original health utility values in 154 studies [14], whilst Bell and colleagues identified 949 health utility values in 228 studies [15]. More recently, systematic reviews of utility values have been reported for a number of specific health states or population groups, for example liver disease [16], neuropathic pain [17], Alzheimer's disease [18], unipolar depression [19], colorectal cancer [20], HIV/AIDS [21], breast cancer [22], type II diabetes [23], surgical site infection [24], Crohn's disease and ulcerative colitis [25], Chronic Obstructive Pulmonary Disease [26], and childhood populations [27]. Methods guidance by some HTA agencies that recommend formal systematic reviews of parameter values for decision-analytic modelling based economic evaluations is likely to increase the number of systematic reviews of health state utility values undertaken [2, 3, 6]. Beyond the context of economic evaluation, evidence from systematic reviews of health state utility values can be used to inform estimates of health burden of disease [28].

There is also growing interest in meta-analytic methods that pool health state utility values collected across a number of studies. Although still relatively rare, these methods generate more precise estimates of the measure of interest, and estimates of uncertainty surrounding those values. Recourse to published meta-analyses of health state utility values should reduce the burden on cost-effectiveness modellers seeking a common source of values for economic evaluations across a clinical area or targeting a specific population group [29].

Building on previous guidance documents [30-32], the purpose of this paper is to provide a practical guide on methods for conducting a systematic review and meta-analysis of health state utility values. The focus is on describing within a single document the possible stages that should be followed in the systematic review and meta-analysis processes in order to enhance transparency, consistency and robustness of methods across studies that synthesise health state utility values.

2. Systematic review methods

The process of systematically reviewing health state utility values should involve a number of formal stages, including identifying the appropriate evidence, study selection, data extraction and presentation, and quality and relevance assessment. Readers are also referred to methodological guidance on the iterative review processes to be followed that has been published elsewhere [30-32]. The general principles that should apply to a well conducted systematic review of health state utility values are that: (i) advice is sought at the outset from an information specialist; (ii) methods are pre-specified in a dated and version-controlled protocol with clearly stated objectives, search terms, literature databases and other sources, inclusion/exclusion criteria, study selection methods, recording of reasons for exclusion, methods for dealing with discrepancies, data extraction templates and planned methods of

synthesis, registered with the Prospective Register of Systematic Reviews (PROSPERO); (iii) approval from an ethics committee or written informed consent are sought if individual-level data are accessed; and (iv) the systematic review is reported in accordance with the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines [33]. The work of searching for and reviewing health state utility values is a specialist task that requires at least some training in economics methods.

2.1 Identification of evidence

Development of a search strategy for a systematic review of health state utility values is likely to require extensive piloting, and an assessment of the appropriate balance to be struck between increased sensitivity on the one hand and potential additional yield on the other. The search strategies of published reviews have relied heavily on the development of bespoke combinations of free-text direct valuation method terms, such as ‘standard gamble’ or ‘time trade-off’ or indirect valuation method terms, such as ‘EQ-5D’ or ‘SF-6D’ or ‘AQL-5D’, combined with relevant terms for the health states or population groups of interest. Particular attention is required to ensure that spelling variants (for example, ‘EQ-5D’ and ‘EQ 5D’), abbreviations (for example, ‘Child Health Utility 9 Dimension’ and ‘CHU-9D’) and synonyms (for example, ‘child’ and ‘kid’ and ‘youth’) are fully captured by the search strategy. The thesauri of major search engines such as Medline (MeSH) and Embase (EMTREE) do not provide granulated thesauri terms for common direct valuation methods, such as ‘standard gamble’, or common multi-attribute health status classification systems with preference scores, such as ‘EQ-5D’ [31]. The focus therefore is largely on identifying the appropriate combination of free text search terms.

Searches of health state utility values have commonly targeted major search engines, such as PubMed, Embase and EconLit. Methods specialist databases, such as the NHS Economic

Evaluation Database (NHS EED) [34] and the Cost-effectiveness Analysis Registry [35] are other potential sources of information. A further potential source of health state utility values is topic specialist or field databases. For example, a recent systematic review of childhood health utilities found that 15 of the 272 eligible studies were only identified through a topic specialist database, namely the Paediatric Economic Database Evaluation (PEDE) Project [27]. Supplementary search strategies include manual reference searching of bibliographies, contacts with experts in the field, citation searching and author searching. A further approach to identifying relevant utility values, which has not been widely applied, is to conduct targeted condition-specific searches of databases of randomised controlled trials (e.g. www.clinicaltrials.gov) and submissions to HTA agencies in order to identify studies that included utility measures as secondary outcomes.

A detailed example of a search strategy developed for the purposes of a systematic review of health state utility values is provided in Appendix A for illustrative purposes.

2.2 Study selection

The stages involved in selecting studies for inclusion in a systematic review of health state utility values are broadly analogous with those followed for systematic reviews of clinical effects, and can be broadly summarised as: (i) examining titles and abstracts to remove obviously irrelevant reports; (ii) retrieving full texts of potentially relevant reports; (iii) linking together multiple reports of the same study; (iv) examining full-text reports for compliance of studies with study eligibility criteria; (v) corresponding with study authors, where appropriate, to clarify study eligibility; and (vi) making final decisions on study inclusion before proceeding to data extraction and assessment [36]. It is recommended that assessments of study eligibility are conducted by at least two people, independently, with disagreements resolved by consensus or the involvement of a third reviewer. There are, in

addition, distinctive features of systematic reviews of health state utility values that merit particular attention when selecting studies. Utility values may be estimated in the context of a number of alternative study designs, including stand-alone preference elicitation studies, randomised controlled trials and various forms of economic evaluation. In the context of some study designs, for example decision-analytic modelling based economic evaluations, it may not be clear at the title and abstract stage whether the authors conducted primary research to estimate utility values or relied purely on secondary data. It may therefore be advisable to be inclusive at the screening stage(s) despite the increased workload likely to be entailed. In addition, justification should be provided for the inclusion of values derived from visual analogue scales that arguably lack a theoretical basis for QALY construction [37].

2.3 Data extraction and presentation

Following the selection of studies for inclusion in a systematic review, data on each study are usually extracted and entered onto a pre-piloted proforma. This process aids the narrative presentation of results, but has other benefits, including providing a final filter for identifying candidate studies that don't meet the inclusion criteria of the review, and identifying data that may be required to modify health state utility values for application in a particular decision model [31]. Good practice guidance for systematic reviews suggests that the data extraction process should be conducted by two independent researchers, followed by a reconciliation process [36]. Available proformas from published reviews have varied in the volume and breath of information that is extracted from individual reports. The type of data extracted should ultimately be guided by the planned presentation of results and the planned analyses. The proforma applied within the systematic review of childhood health utilities is provided in Appendix B for illustrative purposes.

It is preferable to extract alternative descriptive statistics for utility values (for example, means, standard deviations, medians, inter-quartile ranges) where these are available so that analysts can select the preferred statistics for their applications, or transform the data where required. If the study reports utility values for several populations, or sub-groups within one study population, it is preferable to also extract values for each group, thereby enhancing potential applications by analysts.

2.4 Quality and relevance assessment

Critical assessments of the quality of conduct and reporting of contributing studies are important, but constrained by the absence of widely-accepted tools. The CONSORT [38], STROBE [39] and CHEERS [40] statements include salient features for studies that generate utility values within the context of randomised controlled trials, observational studies and single study economic evaluations, respectively. However, those features relate mainly to the underpinning vehicles for data collection rather than the characteristics of health utility assessments. Cooper and colleagues have developed a ranking system for studies that derive health state utility values, ranging from direct utility assessments to Delphi panels and expert opinion [41]. More recently, Papaioannou and colleagues have described key criteria to consider in the utility assessment process, for example, respondent selection and recruitment, response rates to the instrument used, and levels of missing data and how they were dealt with [31]. Other relevant quality assessment criteria have been published [12, 32]. A particular feature of relevance when assessing the quality of contributing studies is their face validity or empirical validity, for example, whether they generate utility values that vary in an expected direction by level of severity for a specific condition [42]. In the absence of generic tools that encompass all potentially relevant features, it is incumbent on those involved in the review process to describe the quality of contributing studies in holistic terms, drawing where

necessary upon the relevant features of multiple checklists. If the systematic review is to be used to inform a pre-specified policy decision, a final stage in the review process should involve a separate assessment of the relevance of contributing studies to the requirements of the local agency considering the evidence [31]. This might include, for example, an assessment of whether contributing studies applied the agency's preferred utility measure or derived values from the preferred population group.

3. Meta-analysis and meta-regression of utility values

When there are multiple estimates of health utilities for a given health state or condition, and if they are sufficiently homogenous, the estimates can be synthesised to provide a pooled estimate of the utility value, which can then be used to populate decision models in economic evaluation or inform broader policy questions [43]. A conservative approach to the inclusion of studies in such a synthesis is advised, bearing in mind the potential use of derived estimates to inform health economic models addressing specific decision problems in specific clinical areas. Health states for different diseases or conditions may share characteristics (e.g. loss of mobility), but have potentially radically different health-related quality of life consequences. In such cases, pooled utility values are unlikely to be appropriate. It may be possible to allow for exceptions where the populations or conditions are broadly similar, or it can be argued on clinical grounds that a systematic relationship exists between states so that their utilities can be jointly estimated allowing for a parameter that defines this systematic relationship. In such cases, pooling of utility values may be appropriate. However, a strong justification should be provided for doing so, and the impact of more restrictive assumptions on utility values should be presented alongside the broader estimates if at all possible.

This section introduces three broad approaches that have been employed in the literature to synthesise comparable utility values: (i) fixed-effect meta-analysis; (ii) random-effects meta-analysis; and (iii) mixed-effects meta-regression. Each approach is illustrated by a synthesis of utility values for a hypothetical decision problem evaluating a new oncology drug that improves survival of paediatric cancer patients across several types of malignancies. CADTH is assumed to be the decision-maker. CADHT accepts both the HUI2 and HUI3 as preference-based multi-attribute utility measures for their reference case despite their differing attributes [2]. It is assumed that CADHT only accepts HUI2 and HUI3 utility values that reflect preferences of the Canadian population over health states. In the absence of any formal guidance by CADTH, utilities measured by different respondent types (e.g. self vs. proxies), administration modes (e.g. self vs. interviewer-administered) and in different years are assumed to be comparable for illustrative purposes. The STATA and R codes used for the alternative approaches to synthesis are provided in Table 1. In addition, the reader is referred to Table 2, which summarises the characteristics of 19 studies identified by a structured review as having synthesised utility values, and which therefore provide a broader context to the methods that are described [16, 17, 19-22, 27, 44-55].

3.1 Fixed-effect meta-analysis

Fixed-effect meta-analysis assumes that utility values are drawn from the same underlying population and that variation is due purely to random error [56]. Figure 1 displays the result of a fixed-effect meta-analysis of 22 mean HUI2 and HUI3 utilities for paediatric cancer survivors, valued using Canadian tariffs, using data from Kwon and colleagues [27]. The exact STATA command using the ‘metan’ package and R command using the ‘metafor’ package is outlined in Table 1. By default, STATA uses inverse-variance weighted fixed-effect meta-analysis. The fixed-effect estimate of mean HUI2/HUI3 utility scores for

paediatric cancer survivors is 0.93 (95% confidence interval: [0.927, 0.936]). There were 16 utility scores published after 2000 and 6 in 2000 or earlier. Although utility scores across different years of publication are seen as comparable in this decision context, separate meta-analyses enable the comparability assumption by year to be tested visually. Heterogeneity between the two groups is statistically insignificant ($P=0.739$) and both groups have a fixed-effect estimate of 0.93. STATA presents Cochran's Q statistics [57] and degrees of freedom for test of heterogeneity, as well as an I^2 statistic where a large I^2 statistic suggests that the level of heterogeneity not attributable to random error is large [58]. From the STATA output we conclude that there is significant variability between individual mean utilities overall ($Q=316.97$; $P<0.001$; $I^2=93.4\%$) and that there is greater variability between mean utilities published after 2000 ($Q=309.48$; $P<0.001$; $I^2=95.2\%$) than those published in 2000 or earlier ($Q=7.38$; $P=0.194$; $I^2=32.2\%$).

3.2 Random-effects meta-analysis

Random-effects meta-analysis relaxes the strong assumption that variation between samples and studies is due solely to random error around the true underlying value. The latter is now assumed to vary across samples and studies. Figure 2 displays the results of a random-effects meta-analysis on the same 22 mean HUI2 and HUI3 utilities for paediatric cancer survivors, this time grouped by HUI2 or HUI3. STATA uses the DerSimonian and Laird method by default [59] and each mean utility value is weighted by the inverse of variance. The τ^2 statistic (reported in non-graph STATA output: $\tau^2=0.0018$ for overall; $\tau^2=0.0049$ for HUI3; $\tau^2=0.009$ for HUI2) estimates the level of between-study variation in underlying utility values. When grouped by HUI2 or HUI3, the I^2 statistics suggest significant heterogeneity between samples in both groups ($I^2=95.9\%$ for HUI3; $I^2=89.1\%$ for HUI2).

Random-effects estimates of mean HUI2 utility values are larger than those of mean HUI3 utility values under both fixed-effect and random-effects meta-analyses (0.92 vs. 0.86 for random-effects and 0.94 vs. 0.90 for fixed-effect). Whether this difference is statistically significant, controlling for other covariates, can be investigated using the meta-regression approach described below. For now, under the assumption in this decision context that HUI2 and HUI3 scores are comparable, the relevant result from Figure 2 is the random-effects estimate of a mean utility value of 0.91 (95% confidence interval, [0.89, 0.93]) for paediatric cancer survivors, combined across HUI2 and HUI3.

Random-effects models can be used to generate ‘mean’ or ‘predictive’ estimates [60, 61]. The former represents the estimated pooled value, whereas the latter represents what might be expected for a new study or population. As seen from Figure 2, the predictive interval [0.82, 1.00] is wider than the confidence interval [0.89, 0.93] for the overall mean estimate (while the predictive interval for HUI3 [0.67, 1.06] crosses the feasible range for utility score), reflecting the additional uncertainty from study/population heterogeneity. Deciding which of these two estimates is relevant depends on the use to which the estimate is to be put. Therefore, studies should preferably report both mean and predictive estimates.

3.3 Mixed-effects meta-regression

Meta-regression is an augmentation of fixed-effect or random-effects meta-analysis with covariates that can partly explain the heterogeneity between samples and studies. Meta-regression with random-effects components is known as mixed-effects meta-regression since heterogeneity is accounted for by a mixture of random-effects and covariates. The mixed-effects approach typically adopts a hierarchical linear structure as follows:

$$U_{ijk} = \beta_0 + \sum_l \beta_l x_{lijk} + \gamma_i + \theta_{ij} + \varepsilon_{ijk}$$

where U_{ijk} is the (weighted) mean of the k^{th} utility of the j^{th} group (defined by combination of covariates) of study i , x_{ijk} the covariates, γ_i the random-effects component of study i , θ_{ij} the random-effects component of j^{th} group of study i , and ε_{ijk} the random error. γ_i , θ_{ij} , and ε_{ijk} are residual variations unexplained by the covariates attributed to different levels within a multilevel structure [55, 62]. The mean utilities are typically weighted by the inverse of variance and clustered by group and study. The study-specific random effects capture the unobservable characteristics of the study (e.g., method of participant recruitment). The studies can also be weighted by their relative ‘importance’, proxied by total number of respondents [21, 27] or total variance (i.e. the sum of all standard errors of mean utilities) [20].

Table 3 shows the results of a mixed-effects meta-regression using the same 22 mean HUI2 and HUI3 utility values for paediatric cancer survivors described above. The explanatory variables all take a binary form, including for the valuation method (reference HUI3 vs. HUI2), respondent type (reference self-response vs. proxy-response), administration mode (reference self-administered vs. interview-administered) and year of publication (reference after 2000 vs. 2000 or earlier).

The constant term in Table 3 suggests that the mean utility estimate for the reference case (i.e. mean HUI3 utility value for paediatric cancer survivors measured using child responses through a self-administered questionnaire) is 0.870 (95% confidence interval: [0.786, 0.953]). The weighted average (using the inverse of prediction standard error as weights) of the mean utilities predicted by the model is 0.937 (95% confidence interval: [0.905, 0.969]). This estimate is higher than the fixed-effect estimate of 0.93 and the random-effects estimate of 0.91.

The advantage of the meta-regression prediction is that it accounts for the statistically significant effects exerted by covariates (in this case, administration mode and year of

publication) not considered in fixed-effect and random-effects meta-analyses. Another advantage is that it allows examination of comparability assumptions, which cannot be done accurately using meta-analysis approaches. In fact, relying on the latter approaches could result in misleading conclusions. For example, Figure 1 suggests that year of publication accounts for an insignificant amount of heterogeneity between mean utilities. Table 3, by contrast, suggests that year of publication exerts a statistically significant effect (coefficient: -0.115; 95% confidence interval [-0.194, -0.036], suggesting earlier published utilities are on average lower than later utilities) once other covariates are held fixed. This questions the initial assumption of treating utility estimates across disparate publication years as comparable. The analyst may therefore, in this decision context, consider including only recently published estimates in the decision model. Another example is the effect of valuation method. Figure 2 suggests that HUI2 utilities are significantly higher than HUI3 utilities under both fixed-effect and random-effects approaches. By contrast, although the direction of coefficient on HUI3 vs. HUI2 in Table 2 (0.043) supports this finding, the coefficient is statistically insignificant ($P=0.068$) once other covariates are held fixed. This supports the initial assumption that HUI2 and HUI3 can be treated as comparable measures for paediatric cancer survivors.

It should be noted that the mixed-effect meta-regression model in Table 3 tests four hypotheses simultaneously. This could result in type I errors (false-positives) where significant findings are driven by natural variation across multiple subgroups rather than non-random effects of the explanatory variables [63]. This problem is particularly acute in meta-regressions where there is significant between-study and between-group heterogeneity [64]. Bonferroni correction [65] is a simple but conservative way to address this issue by lowering the nominal significance level. If the significance level is $1 - \alpha$ then only the results with P -values less than α/k where k is the number of tests are interpreted as significant. Hence, with

$\alpha = 0.05$ and $k = 4$, the relevant P -value is 0.0125. Even so, in our example, both administration mode ($P < 0.001$) and year of publication ($P = 0.004$) are still associated with significant effects on utility values.

4. Methodological considerations for meta-analysis and meta-regression

Although meta-analysis and meta-regression improve our estimates of health utility values for a given health state or condition by making use of relevant sources of information across multiple studies and samples, analysts should be acquainted with several methodological considerations and limitations associated with their application.

4.1 Selection of comparable utilities for synthesis

It is important to limit synthesis to ‘comparable’ utility estimates, even in meta-regression where confounding factors are controlled for. In the examples above, for instance, utilities valued using different country-specific tariffs are *not* deemed comparable while HUI2 and HUI3 utilities (derived using Canadian tariffs) *are* deemed comparable. There is no fixed rule determining the bounds of comparability. Ideally, the acceptable level of comparability between health states (e.g. utility estimates for survivors of different types of cancer) should be defined by the decision problem and health states populating the decision model, while the acceptable level of comparability for methodological factors should ideally be guided by the decision-making body (e.g. the preferred utility measurement approach and source of valuation).

It is often unclear whether or not multiple utility estimates are reasonably comparable, especially regarding methodological factors such as respondent type, administration mode

and closely-related valuation methods (e.g. HUI2 vs. HUI3). A potential approach is to conduct a meta-regression as in Section 3.3 to identify methodological factors that exert statically significant effects on utility estimates. Analysts can then either: (i) use the predicted values of the meta-regression model that accounts for the independent effects; or (ii) conduct separate meta-analyses for sub-groups defined by the significant methodological factors. In addition, published meta-regression studies often explore statistical significance of clinical and methodological factors in meta-regression models [16, 17, 20-22, 27, 46, 47, 50, 52-55]; their results could hence serve as a guide on selecting a preliminary set of covariates for regression. However, caution is required when interpreting previous meta-regression results since it is unclear whether they are generalisable to different contexts, while statistically significant coefficients may be due to ecological fallacy (see Section 4.3), small sample sizes or multiple comparisons. Selection of covariates could also be informed by Akaike's information criterion for nested models [49].

4.2 Insufficient information from published studies

One major limitation of secondary evidence synthesis is its reliance on limited or poorly reported published data in primary utility assessment studies. The analyst's response in this circumstance should depend on the nature of insufficient information.

4.2.1 Insufficient information on health states

Reporting of clinical details within utility assessments may be poor or idiosyncratic, resulting in non-uniform categorisation of health states across studies. This risks synthesising information on qualitatively different health states. Analysts should refer to external guidelines and epidemiological data to either clarify or re-classify health states to obtain

internal comparability with other samples. For example, Tengs and Lin [21] and Tran and colleagues [54] found that most but not all primary utility studies in the HIV/AIDS context classified HIV/AIDS stages as asymptomatic HIV, symptomatic HIV, and AIDS. Those that did not only reported CD4 counts and specific HIV/AIDS symptoms. Therefore, both meta-analyses utilise epidemiological data to impute HIV/AIDS stage from CD4 counts and/or specific symptoms. Other commonly missing clinical details include disease severity, treatment regime for chronic diseases and time since diagnosis and intervention, as well as demographic factors such as age and gender.

4.2.2 Insufficient information on methodological factors

Paucity of information is also often a feature for methodological factors. Analysts should pay close attention to the description of study features in primary studies to extract information on, for example, administration mode and respondent type since these are often not explicitly stated. Moreover, primary studies may use and report several multi-attribute utility measures, all of which can potentially be acceptable in the decision context, but not detail which of the measures have psychometric properties best suited to the given health state or condition.

Another component of commonly missing information relates to the population and valuation method by which tariffs for preference-based multi-attribute utility measures are derived.

Primary studies frequently only provide a reference to the tariff valuation study, leaving the analyst to chase the reference to obtain relevant details.

4.2.3 Synthesis of samples with insufficient information

The analyst may face a trade-off on whether to include samples lacking in clinical or methodological details. Inclusion risks synthesis of potentially incomparable estimates whilst exclusion limits the sample size. Again, the most reasonable approach would depend on the

decision problem and context. If the decision model requires utility estimates for a specific level of disease severity, then samples with ambiguous severity descriptions should be excluded. If the decision-making body does not have a preferred respondent type or administration mode, including samples without clear description of these factors may be acceptable. For some health states (e.g., childhood health states or health states related to cognitive impairment), respondent type and administration mode may exert a significant effect on questionnaire response. In such cases, samples without clear description of these factors should be excluded.

4.3 Use of aggregate outcomes

Most primary studies assessing utilities only report aggregate outcomes such as sample mean or median utilities and measures of variability for the point estimates. There is a risk of ecological fallacy where the nature of associations between covariates at the aggregate or population level is different from those at the patient or individual level [66]. For example, the analyst may be interested in heterogeneity in utility score by gender. However, primary studies often do not report gender-specific (all-male or all-female) sub-samples and only report aggregate utility values for gender-mixed samples. In these cases, including proportion of males as a covariate in meta-regression may produce a very different nature of association than individual-level analysis.

4.4 Measure of uncertainty for utility estimates

Synthesis of mean utilities should use the same measure of uncertainty to weight individual mean utilities. The most widely used weight is the inverse variance weight constructed from the sample standard error. Missing standard errors can be imputed based on sample characteristics using information from comparable samples within other primary studies that

do report standard errors [53]. A cruder method is to assign the average of standard errors for comparable samples [47]. In some cases, the standard deviation is reported but sample size is missing. Here a crude method is to assign a sample size of one, or if the utility estimate was derived through expert opinion, the number of experts could be used for the sample size.

4.5 Use of median utilities

Median utilities and their interquartile ranges are in themselves informative central point estimates and measures of variability. These statistics should therefore be extracted from primary studies alongside means, standard errors and ranges. Additional considerations are required, however, before median values are synthesised with mean values. Motivation for this may be to form an adequate sample size and prevent loss of information [22, 50, 55]. Previous meta-analyses have justified this in various ways. Sturza inspected the skewness of the interquartile range, and because skewness was small, median utilities were treated as means, and the average of sample variances reported by other samples was used as weights for the medians [50]. Peasgood and colleagues mapped median utilities onto mean values according to an estimated association between median and mean utilities reported by the same study [22].

4.6 Synthesis of longitudinal utility estimates

Primary studies reporting health state utility values can be derived using various study designs, including cross-sectional observational studies, longitudinal observational studies and prospective randomised trials. The latter two designs can generate information on longitudinal trajectory of utility estimates. For example, Han and colleagues assess utilities

for two sample of adolescents admitted to intensive care for injuries of various types (one sample had post-injury depression and the other did not) using the Quality of Well-Being (QWB) measure at 3-month, 6-month, 12-month, 18-month and 24-month follow-up points [67]. The analyst could conceivably utilise all ten utility estimates by, for example, including follow-up time as a covariate in a meta-regression model. However, this approach ignores the correlation of utility estimates across time [53], which will not be completely captured by within-study clustering within hierarchical linear models. Moreover, the time-utility relationship is unlikely to be linear [54]. Therefore, as far as possible, analysts should avoid synthesising utility estimates across different time points. Peasgood and colleagues address this in their analysis by conducting separate syntheses of utilities for osteoporosis-related fractures by category of time since fracture [49].

4.7 Intervention effects on utility estimates

Closely associated with longitudinal measurement of utility values is the comparison of pre- and post-intervention utility estimates in longitudinal observational and experimental primary studies. For example, in a randomised controlled trial investigating the cost-effectiveness of a combined physical exercise and psychosocial training intervention for paediatric cancer survivors, Braam and colleagues assess utilities of participants in both intervention and control groups using the EQ-5D-Youth (EQ-5D-Y) measure at pre-treatment baseline and 12-month follow-up points [68]. Post-treatment utility estimates offer important information on treatment effects that last for a period of time and have health-related quality-of-life implications. However, analysts should be cautious about synthesising utility estimates across treatment status and then including treatment as a covariate in a meta-regression model, as has been done in many previous analyses (e.g. [52, 54]). First, it may be difficult to

disentangle between treatment effects and age-related changes in utility over time. Second, any estimate of treatment effect using aggregate outcomes would differ from individual-level estimates [47]. Moreover, endogeneity may be more acute when aggregate outcomes across multiple studies are used, since health status and choice of treatment modality may be closely correlated [48].

Therefore, as far as possible, analysts should refrain from synthesising utilities with heterogeneous treatment status and post-treatment time profiles, although this can be difficult if primary studies do not offer adequate information, especially for chronic disease health states. In the context of a longitudinal experimental study, such as that by Braam and colleagues [68], analysts could treat baseline utilities from both intervention and control groups as estimates of utility for paediatric cancer survivors not confounded by any post-cancer treatment. These utilities could potentially be synthesised with comparable utilities for paediatric cancer survivors from other studies similarly not confounded by any further treatment. The difference in mean utilities at the 12-month follow-up between the treatment and control groups within the study by Braam and colleagues could serve as an estimate of intervention effect. If there are other studies investigating the health utility effects of the same intervention, the treatment effects could be synthesised and the resulting estimate could be used to inform parameter values for an intervention-related state within a decision model.

4.8 Model validity

Many studies conducting meta-regression find it desirable to test the fit of their models. The R^2 value has been used [20], as has been Akaike's information criterion [16], in an analogous way to the use of I^2 and χ^2 statistics for fixed-effect and random-effects meta-analysis. Some studies have plotted predicted values against observed values [20]. Another measure of fit is

the proportion of predicted values above 1.0 [27]. The use of a generalised linear model with non-identity link (e.g., logit link) is a way of setting a bound for utility to the 0-1 range [47, 20]. However, this makes interpretation of coefficients difficult [20], whilst some research suggests that non-linear models may not produce better fit than linear models [47, 69].

5. Conclusion

Economic analysts are likely to increasingly rely on systematic reviews and meta-analyses of health state utility values to inform the parameter inputs of decision-analytic modelling based economic evaluations. Beyond the context of economic evaluation, evidence from systematic reviews and meta-analyses of health state utility values can be used to inform broader health policy decisions. This paper provides practical guidance on how to conduct a systematic review and meta-analysis of health state utility values. The paper outlines a number of stages in conducting a systematic review of health state utility values. It further describes three broad approaches that have been employed in the literature to synthesise multiple estimates of health utilities for a given health state or condition, namely fixed-effect meta-analysis, random-effects meta-analysis, and mixed-effects meta-regression. Each approach is illustrated by a synthesis of utility values for a hypothetical decision problem, and software code is provided. The paper highlights a number of methodological issues pertinent to the conduct of meta-analysis or meta-regression. Approaches for addressing methodological challenges are presented and issues of methodological debate are highlighted.

Avenues for further research clearly arise from the material in this guidance document. The main evidence gap is in identifying a preferred approach for pooling health state utility values collected across a number of studies and generating estimates of uncertainty surrounding those values. The available approaches may best be viewed as complements

rather than competing alternatives. Selection of one approach should be informed by a combination of clinical and statistical judgment, and an assessment of the use to which the outputs will be put. Bayesian methods are increasingly being recommended by HTA agencies for use in meta-analysis of treatment effects, because they align well with decision modelling analyses, and provide additional benefits when constructing multi-level models [70].

However their use in the context of utility value synthesis is sparse, and further research to explore the value of Bayesian meta-analytic methods in this setting is clearly required.

Acknowledgments

We are grateful to departmental colleagues for their comments on the paper and suggestions provided. SP receives financial support as a National Institute for Health Research Senior Investigator. No specific funding was obtained to produce this paper. The authors do not have any conflicts of interest to declare.

Author's contributions

All authors contributed to the conception, design and drafting of the paper. All authors reviewed and approved the final version of the paper. SP is the guarantor of the overall content.

References

1. PBAC (Pharmaceutical Benefits Advisory Committee). Guidelines for preparing submissions to the pharmaceutical benefits advisory committee. Australia: Australian Government Department of Health; 2013.
2. CADTH (Canadian Agency for Drugs and Technologies in Health). Guidelines for the economic evaluation of health technologies. Canada: CADTH; 2006.
3. HAS (Haute Autorité de Santé). Choices in methods for economic evaluation. France: HAS; 2012.
4. CVZ (College voor zorgverzekeringen). Guidelines for pharmacoeconomic research: evaluation and actualisation. Diemen: CVZ; 2006.
5. CatSalut. Guia I Recomanacions Per A La Realització I Presentació D'avaluacions Econòmiques I Anàlisis D'impacte Pressupostari De Medicaments En L'àmbit Del Catsalut. Catalonia: CatSalut; 2014.
6. NICE (National Institute of Health and Care Excellence). Guide to the methods of technology appraisal. London: NICE; 2013.
7. SMC (Scottish Medicines Consortium). Guidance to manufacturers for completion of new product assessment form (NPAF). Scotland: NHS Scotland; 2016.
8. Brooks R. EuroQol: The current state of play. *Health Policy*. 1996;37(1):53-72.
9. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. *Health Utilities Index*. *Pharmacoeconomics*. 1995;7(6):490-502.
10. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ*. 2002;21(2):271-92.
11. Brazier J, Ratcliffe J, Salomon J, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2007.

12. Wolowacz SE, Briggs A, Belozeroff V, Clarke P, Doward L, Goeree R et al. Estimating health-state utility for economic models in clinical studies: An ISPOR Good Research Practices Task Force Report. *Value Health*. 2016;19(6):704-19.
13. Sculpher MJ, Claxton K, Drummond M, McCabe C. Whither trial-based economic evaluation for health care decision making? *Health Econ*. 2006;15(7):677-87.
14. Tengs TO, Wallace A. One thousand health-related quality-of-life estimates. *Med Care*. 2000;38(6):583-637.
15. Bell CM, Chapman RH, Stone PW, Sandberg EA, Neumann PJ. An off-the-shelf help list: a comprehensive catalog of preference scores from published cost-utility analyses. *Med Decis Making*. 2001;21(4):288-94.
16. McLernon DJ, Dillon J, Donnan PT. Health-state utilities in liver disease: a systematic review. *Med Decis Making*. 2008;28(4):582-92.
17. Doth AH, Hansson PT, Jensen MP, Taylor RS. The burden of neuropathic pain: a systematic review and meta-analysis of health utilities. *Pain*. 2010;149(2):338-44.
18. Shearer J, Green C, Ritchie CW, Zajicek JP. Health state values for use in the economic evaluation of treatments for Alzheimer's disease. *Drugs Aging*. 2012;29(1):31-43.
19. Mohiuddin S, Payne K. Utility values for adults with unipolar depression: Systematic review and meta-analysis. *Med Decis Making*. 2014;34(5):666-85.
20. Djalalov S, Rabeneck L, Tomlinson G, Bremner KE, Hilsden R, Hoch JS. A review and meta-analysis of colorectal cancer utilities. *Med Decis Making*. 2014;34(6):809-18.
21. Tengs TO, Lin TH. A meta-analysis of utility estimates for HIV/AIDS. *Med Decis Making*. 2002;22(6):475-81.
22. Peasgood T, Ward SE, Brazier J. Health-state utility values in breast cancer. *Expert Rev Pharmacoecon Outcomes Res*. 2010;10(5):553-66.

23. Beaudet A, Clegg J, Thuresson PO, Lloyd A, McEwan P. Review of utility values for economic modeling in type 2 diabetes. *Value Health*. 2014;17(4):462-70.
24. Gheorghe A, Moran G, Duffy H, Roberts T, Pinkney T, Calvert M. Health utility values associated with surgical site infection: A systematic review. *Value Health*. 2015;18(8):1126-37.
25. Malinowski KP, Kawalec P. Health utility of patients with Crohn's disease and ulcerative colitis: a systematic review and meta-analysis. *Expert Rev Pharmacoecon Outcomes Res*. 2016;16(4):441-53.
26. Moayeri F, Hsueh YS, Clarke P, Hua X, Dunt D. Health state utility value in Chronic Obstructive Pulmonary Disease (COPD); The challenge of heterogeneity: A systematic review and meta-analysis. *COPD*. 2016;13(3):380-98.
27. Kwon J, Kim SW, Ungar WJ, Tsiplova K, Madan J, Petrou S. A systematic review and meta-analysis of childhood health utilities. *Med Decis Making*. 2017:272989X17732990.
28. Brown DS, Trogdon JG, Ekwueme DU, Chamiec-Case L, Guy GP, Jr., Tangka FK et al. Health state utility impact of breast cancer in U.S. women aged 18-44 years. *Am J Prev Med*. 2016;50(2):255-61.
29. Karnon J. Health state utility values for cost-effectiveness models. *Pharmacoeconomics*. 2017. doi:10.1007/s40273-017-0537-x
30. Papaioannou D, Brazier J, Paisley S. NICE DSU Technical Support Document 9: The identification, review and synthesis of health state utility values from the literature. NICE Decision Support Unit Technical Support Documents. London; 2010.
31. Papaioannou D, Brazier J, Paisley S. Systematic searching and selection of health state utility values from the literature. *Value Health*. 2013;16(4):686-95.

32. Ara R, Brazier J, Peasgood T, Paisley S. The identification, review and synthesis of health state utility values from the literature. *Pharmacoeconomics*. 2017;35(Suppl 1):43-55.
33. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*. 2009;339:b2535.
34. NHS Economic Evaluation Database [database on the Internet]. Centre for Reviews and Dissemination. 2010. Available from: <http://www.york.ac.uk/inst/crd>. Accessed: December 2017.
35. CEA Registry [database on the Internet]. Tufts Medical Centre. 2010. Accessed: December 2017.
36. Higgins JPT. Cochrane handbook for systematic reviews of interventions. Available from www.handbook.cochrane.org: The Cochrane Collaboration; 2011.
37. Parkin D, Devlin N. Is there a case for using visual analogue scale valuations in cost-utility analysis? *Health Econ*. 2006;15(7):653-64.
38. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c332.
39. Vandembroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Ann Intern Med*. 2007;147(8):W163-94.
40. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)-explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. *Value Health*. 2013;16(2):231-50.

41. Cooper N, Coyle D, Abrams K, Mugford M, Sutton A. Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. *J Health Serv Res Policy*. 2005;10(4):245-50.
42. Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. *Health Econ*. 1999;8(1):41-51.
43. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices-Modeling Studies. *Value Health*. 2003;6(1):9-17.
44. Cheng AK, Niparko JK. Cost-utility of the cochlear implant in adults: a meta-analysis. *Arch Otolaryngol Head Neck Surg*. 1999;125(11):1214-8.
45. Post PN, Stiggelbout AM, Wakker PP. The utility of health states after stroke: a systematic review of the literature. *Stroke*. 2001;32(6):1425-9.
46. Tengs TO, Lin TH. A meta-analysis of quality-of-life estimates for stroke. *Pharmacoeconomics*. 2003;21(3):191-200.
47. Bremner KE, Chong CA, Tomlinson G, Alibhai SM, Krahn MD. A review and meta-analysis of prostate cancer utilities. *Med Decis Making*. 2007;27(3):288-98.
48. Liem YS, Bosch JL, Hunink MG. Preference-based quality of life of patients on renal replacement therapy: a systematic review and meta-analysis. *Value Health*. 2008;11(4):733-41..
49. Peasgood T, Herrmann K, Kanis JA, Brazier JE. An updated systematic review of Health State Utility Values for osteoporosis related conditions. *Osteoporos Int*. 2009;20(6):853-68.
50. Sturza J. A review and meta-analysis of utility values for lung cancer. *Med Decis Making*. 2010;30(6):685-93.

51. Lung TW, Hayes AJ, Hayen A, Farmer A, Clarke PM. A meta-analysis of health state valuations for people with diabetes: explaining the variation across methods and implications for economic evaluation. *Qual Life Res.* 2011;20(10):1669-78.
52. Wyld M, Morton RL, Hayen A, Howard K, Webster AC. A systematic review and meta-analysis of utility-based quality of life in chronic kidney disease treatments. *PLoS Med.* 2012;9(9):e1001307.
53. Si L, Winzenberg TM, de Graaff B, Palmer AJ. A systematic review and meta-analysis of utility-based quality of life for osteoporosis-related conditions. *Osteoporos Int.* 2014;25(8):1987-97.
54. Tran BX, Nguyen LH, Ohinmaa A, Maher RM, Nong VM, Latkin CA. Longitudinal and cross sectional assessments of health utility in adults with HIV/AIDS: a systematic review and meta-analysis. *BMC Health Serv Res.* 2015;15:7.
55. Sampson CJ, Tosh JC, Cheyne CP, Broadbent D, James M. Health state utility values for diabetic retinopathy: protocol for a systematic review and meta-analysis. *Syst Rev.* 2015;4:15.
56. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for meta-analysis in medical research.* Chichester: John Wiley & Sons; 2000.
57. Cochran WG. The comparison of percentages in matched samples. *Biometrika.* 1950;37(3-4):256-66.
58. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med.* 2002;21(11):1539-58.
59. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials.* 1986;7(3):177-88.
60. Ades AE, Lu G, Higgins JP. The interpretation of random-effects meta-analysis in decision models. *Med Decis Making.* 2005;25(6):646-54.

61. Welton NJ, Soares MO, Palmer S, Ades AE, Harrison D, Shankar-Hari M et al. Accounting for heterogeneity in relative treatment effects for use in cost-effectiveness models and value-of-information analyses. *Med Decis Making*. 2015;35(5):608-21.
62. Kalaian HA, Raudenbush, S.W. A multivariate mixed linear model for meta-analysis. *Psychological Methods*. 1996;1(3):227-35.
63. Bland M. An introduction to medical statistics, 3rd edition. USA: Oxford University Press; 2000.
64. Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Stat Med*. 2004;23(11):1663-82.
65. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Stat Med*. 2014;33(11):1946-78.
66. Dias S, Sutton AJ, Welton NJ, Ades AE. Heterogeneity: Subgroups, Meta-Regression, Bias And Bias-Adjustment. NICE Decision Support Unit Technical Support Documents. London; 2012.
67. Han PP, Holbrook TL, Sise MJ, Sack DI, Sise CB, Hoyt DB et al. Postinjury depression is a serious complication in adolescents after major trauma: injury severity and injury-event factors predict depression and long-term quality of life deficits. *J Trauma*. 2011;70(4):923-30.
68. Braam KI, van Dijk-Lokkart EM, van Dongen JM, van Litsenburg RRL, Takken T, Huisman J et al. Cost-effectiveness of a combined physical exercise and psychosocial training intervention for children with cancer: Results from the quality of life in motion study. *Eur J Cancer Care (Engl)*. 2017;26(6).
69. Pullenayegum EM, Tarride JE, Xie F, Goeree R, Gerstein HC, O'Reilly D. Analysis of health utility data when some subjects attain the upper bound of 1: are Tobit and CLAD models appropriate? *Value Health*. 2010;13(4):487-94.

70. Dias S, Welton NJ, Sutton AJ, Ades AE. A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials. NICE Decision Support Unit Technical Support Documents. London; 2014.

Table 1. STATA and R codes for fixed-effect and random-effects meta-analysis and mixed-effects meta regression

Code	Note
STATA	
ssc install metan (https://www.stata.com/support/faqs/statistics/meta-analysis/)	
STATA fixed-effect meta-analysis	
metan huimean se, by(yearpub) lcols(ref) xlabel(-0.3,1)	“huimean”: mean HUI2 and HUI3 utilities “se”: standard error of mean utilities “yearpub”: year of publication, binary variable “ref”: study reference
STATA random-effects meta-analysis	
metan huimean se, random by(huicat) lcols(ref) xlabel(-0.3,1) second(fixed) rfdist	“random”: specifies random-effects “huicat”: HUI3 vs. HUI2, binary variable “second(fixed)”: run fixed-effect analysis for comparison “rfdist”: prediction interval is given for the random-effects estimate
STATA mixed-effects meta-regression	
xtmixed huimean huicat restype admin yearpub [pw=inversese], ref:, pweight(refwgt)	“restype”: respondent type, binary variable “admin”: administration mode, binary variable “[pw=inversese]”: specifies variable “inversese”, the inverse of sample standard error, as weights for mean utilities “ ref:”: treats study reference as higher level under which mean utilities are clustered “pweight(refwgt)”: weights references by variable “refwgt” which could be sum of sample standard errors
R	
install.packages(“metafor”) (https://cran.r-project.org/web/packages/metafor/index.html) library(metafor)	
R fixed-effect meta-analysis	
fixed <- rma(yi=huimean, cei=se, data=meta, method='FE', subset=(yearpub>2000)	“data=meta”: assumes that data is already in R format “method='FE'”: specifies fixed effect option “subset=()”: specifies subgroup for analysis
forest(fixed, slab=paste(meta\$ref))	“forest()”: draws the forest plot
R random-effects meta-analysis	
random <- rma(yi=huimean, cei=se, data=meta)	
predict(random)	“predict()”: gives predicted estimate and prediction interval
R mixed-effects meta-regression	
install.packages(“lme4”) library(lme4) mixed <- lmer(utilmean ~ valcode+rescode+modecode+yearpub + (1 ref),data=meta, weights=(meta\$utilse), REML=TRUE) summary(mixed)	To our knowledge, LME4 package does not allow weights for both level 1 and level 2 variables. Hence mean utility is weighted but study is not. “REML=TRUE” specifies restricted maximum likelihood estimation technique which produces unbiased coefficient estimates

Table 2. Previous systematic reviews and meta-analyses of health utility values

Reference	Health State/ Condition	Review Period	Statistical Model	Sample Size	Dependent Variable	Pooling Sub- Groups/Covariates
Cheng and Niparko (1999) [44]	Profound deafness and cochlear implant in adults	January 1966 to May 1999	Fixed-effect meta-analysis models	(i) 7 mean utility gains from 6 studies (ii) 9 mean utility losses from 5 studies	For all models: mean utility gain/loss from HUI, VAS and QWB weighted by inverse of sample variance	(i) Pooling of mean utility gains from cochlear implant (ii) Pooling of mean utility losses from profound deafness
Post et al. (2001) [45]	Stroke	Inception to 2000	Fixed-effect meta-analysis models	43 mean utilities from 23 studies	(i) Normalised mean utility from TTO, SG, VAS, HUI and EQ-5D weighted by sample size (ii) Mean utility from EQ-5D scored by authors	(i) Pooling by respondent type (healthy population, patients at risk for stroke, stroke survivors), valuation method and stroke severity (minor, major, unspecified) (ii) Pooling by stroke severity for EQ-5D only
Tengs and Lin (2002) [21]	HIV/AIDS	1985 to 2000	Mixed-effects meta-regression model (HLM)	74 mean/median utilities from 25 studies	Mean/median utility from SG, TTO, QWB, VAS, and expert judgement. Studies weighted by	Disease stage (asymptomatic HIV, symptomatic HIV, AIDS); Valuation method; Respondent type; Lower bound of scale; Upper bound of scale; Year of publication

					number of respondents	
Tengs and Lin (2003) [46]	Stroke	1985 to 2000	Mixed-effects meta-regression model (HLM)	53 mean utilities from 20 studies	Mean utility from VAS, TTO, SG, and expert judgement weighted by sample size	Stroke severity (minor, moderate, major); Valuation method; Respondent type; Upper/lower bound of scale
Bremner et al. (2007) [47]	Prostate cancer	1966 to September 2004	(i) Fixed-effect meta-analysis models (ii) Mixed-effects meta-regression models (HLM)	173 mean utilities from 23 studies	Mean utility from TTO, SG/HUI, VAS/QWB, and expert judgement weighted by product of sample size and inverse of sample variance	(i) Fixed-effect model: Pooling by cancer symptom and severity or by source (respondent type and valuation of hypothetical scenarios) and valuation method (ii) HLM: Cancer stage (metastatic, nonmetastatic, mixed); Cancer symptom; Cancer severity; Source; Valuation method; Upper bound of scale; Study design (CDA vs. utility assessment); Administration mode
Liem et al (2008) [48]	End-stage renal disease; renal replacement therapy	Inception to September 2006	Random-effects meta-analysis models	62 mean utilities from 27 studies	Mean utilities pooled separately for VAS, TTO, SG, EQ-5D and HUI	Pooling by valuation method and treatment modality (peritoneal dialysis,

						hemodialysis, kidney transplant)
McLernon et al. (2008) [16]	Chronic liver disease	1966 to September 2006	Mixed-effects meta-regression model (HLM)	40 mean utilities from 6 studies (restricted to disease states with at least 3 samples)	Mean utility from EQ-5D, VAS, SG mapped from VAS, TTO, SG, HUI2 and HUI3 weighted by inverse of sample variance	Hepatitis C disease states ^a ; Valuation method
Peasgood et al. (2009) [49]	Osteoporosis related conditions	January 2000 to July 2007	Fixed-effect meta-analysis models	Mean utilities from 27 studies	Mean utility from EQ-5D weighted by inverse of sample variance or by sample size	Pooling of EQ-5D index utilities only by osteoporosis-related fractures (pre-fracture, vertebral, hip, wrist, shoulder, multiple) and year since fracture
Doth et al. (2010) [17]	Neuropathic pain	Inception to April 2008	(i) Random-effects meta-analysis models (ii) Mixed-effects meta-regression model (HLM)	(i) 22 mean utilities from 13 studies (ii) Mean utilities (number not stated) from 10 studies	(i) Mean utility from EQ-5D only weighted by inverse of sample variance (ii) Mean utility from EQ-5D, SG, 15D, Global Rating of Health Care (weights not stated)	(i) Random-effects models: Pooling by six neuropathic pain conditions ^b (ii) HLM: Neuropathic pain conditions; Mean age; Sex; Pain duration; Pain severity; Comorbidities; Valuation method

Peasgood et al. (2010) [22]	Breast cancer	Inception to March 2009	Meta-regression using simple, pooled OLS models with standard errors robust to within-study clustering for: (i) Early breast cancer (EBC) (ii) Metastatic breast cancer (MBC)	(i) EBC models: Up to 230 mean utilities from 29 studies ^c (ii) MBC models: Up to 117 mean utilities from 20 studies ^d	(i) EBC models: Mean utility from SG, VAS (worst-best, dead-full) EQ-5D, TTO (top bound full health, not full health, other), and HUI3 (ii) MBC models: Mean utility from VAS (worst-best, dead-full, rescaled to dead-full), EQ-5D, and TTO (top bound full health, not full health) For both (i) and (ii): Separate model excluding VAS and TTO without full health top bound; Weighted by inverse of standard deviation or sample size	(i) EBC models: Surgery type; Nonsurgical treatment type; Time since diagnosis or treatment; Valuation method; Respondent type; Hypothetical scenario valuation. (ii) MBC models: Treatment type; Response to treatment; Side-effects; Valuation method; Respondent type; Hypothetical scenario valuation
Sturza (2010) [50]	Lung cancer	Not stated	Mixed-effects meta-	223 mean/median	Mean ^e utility from SG, VAS, HALex, AQoL, EQ-5D,	Valuation method; Respondent type; Lung cancer severity; Lung cancer

			regression model (HLM)	utilities from 23 studies	TTO, and Expert judgement weighted by inverse of sample variance (sensitivity analysis without weights)	type (metastatic, non-metastatic, mixed/not stated); Lower bound of scale; Upper bound of scale; Year 2005 in sensitivity analysis
Lung et al (2011) [51]	Diabetes and diabetes-related complications	Inception to end of 2009	(i) Random-effects meta-analysis models (ii) Mixed-effects meta-regression model (HLM)	(i) Random-effects models: 66 mean utilities from 45 studies (ii) HLM: 59 mean utilities from 40 studies	Mean utility from EQ-5D, TTO, SG, HUI2, HUI3, and SF-6D weighted by inverse of sample variance	(i) Random-effects models: Pooling across all diabetes conditions and separately by seven diabetes-related complications (ii) HLM: Total number of respondents; Sample mean age; Proportion of males in sample; Valuation method
Wyld et al. (2012) [52]	Chronic kidney disease	Inception to 2009	Mixed-effects meta-regression models (HLM)	For all models: 326 mean utilities from 190 studies	Mean utility from TTO, SG, AQoL, EQ-5D, SF-6D, 15D, HUI2/3, and EQ-5D mapped from SF-36 and SF-12	HLMs: (a) Treatment type (pre-dialysis, dialysis, kidney transplant, or non-dialytic therapy); Valuation method (b) Haemodialysis vs. Peritoneal dialysis; Valuation method (c) Automated vs. ambulatory peritoneal dialysis; Valuation method (d) Proportion of sample

						diabetic; Treatment modality; Valuation method (e) Year of publication; Valuation method
Djalalov et al. (2014) [20]	Colorectal cancer	January 1980 to January 2013	Mixed-effects meta-regression models (HLM) with Bayesian priors for coefficients, intercept, random effects and error term	(i) Baseline HLMs: 157 mean utilities from 26 studies (ii) Supplemental HLM ^f : 351 mean utilities from 26 studies	For all models: Mean utility from TTO, SG, VAS, EQ-5D, and HUI3 weighted by inverse of sample variance. Studies weighted by inverse of the sum of sample variances.	For all models: Cancer site (colorectal, colon, rectal); Cancer stage; Time to or from initial care; Valuation method; Administration mode
Mohiuddin and Payne (2014) [19]	Unipolar depression in adults without significant comorbidity	1946 to 2012	Random-effects meta-analysis models using Bayesian uninformative prior for between-study variance	18 mean utilities from 6 studies ^g	Mean utility from SG or EQ-5D weighted by inverse of sample variance	Pooling by depression severity (mild, moderate, severe) and valuation method (SG, EQ-5D)
Si et al. (2014) [53]	Osteoporosis related conditions	Inception to 2013	(i) Random-effects meta-analysis models	362 mean utilities from 62 studies: 106 mean utilities	Mean utility from EQ-5D, HUI, QWB, SF-6D, RS, VAS, SG, and TTO	(i) Random-effects models: Pooling by osteoporosis-related condition

			(ii) Univariate mixed-effects meta-regression models (HLM) (iii) Multivariate HLMs	for pre-fracture, 89 for hip fracture, 130 for vertebral fracture, and 37 for wrist fracture	weighted by inverse of sample variance	(ii) For each osteoporosis-related condition, separate univariate HLM for: Time since fracture; Age group; Valuation method; Retrospective assessment; Country; Sex; Fracture history (iii) Multivariate HLM for each osteoporosis-related condition: Time since fracture; Age; Valuation method; Retrospective assessment; Sex; Fracture history
Tran et al. (2015) [54]	HIV/AIDS	2000 to February 2014	(i) Mixed-effects meta-regression models (HLM) for cross-sectional data (ii) 1 st and 2 nd order fractional polynomial regression models (FPR) for	(i) HLM: 218 mean utilities from 49 studies (ii) FPR: 99 mean utilities from 14 studies	For all models: Mean utility from EQ-5D, HUI2/3, SF-6D, 15D, SG, TTO, and VAS	(i) HLM: Valuation method; Disease stage (asymptomatic HIV, symptomatic HIV, AIDS); Treatment type; Country setting (developed vs. developing); Year of publication (ii) FPR: Valuation method; Length of ART; Country setting; Year of publication

longitudinal
change in
utility over
duration of
ART

Sampson et al. (2016) [55]	Diabetic retinopathy and maculopathy	Inception to 2015	(i) Fixed-effect meta-analysis models (ii) Univariate mixed-effects meta- regression models (HLM) (iii) Multivariate HLM	For all models: 317 mean/median utilities from 41 studies. Assume that median utilities are normally distributed and treat as mean values, using IQR to estimate standard error	(i) Fixed-effect models: EQ-5D, EQ-5D VAS, HUI3, 15D, SF-6D, SG, and TTO (ii-iii) For HLMs: EQ-5D, EQ-5D VAS, HUI3, 15D, SF-6D, SG, TTO, QWB, and VAS. Mean utilities weighted by product of sample size and inverse of sample variance	(i) Fixed-effect models: Pooling by retinopathy and maculopathy (ROM0) grades and valuation method (ii) Univariate HLMs: ROM0 grades; Valuation method; Year of publication; Study design; Diabetes type; Tariff population; Respondent type, Cooper quality score; Administration mode (iii) Multivariate HLM: ROM0 grades; Valuation method; Tariff population; Respondent type
Kwon et al. (2017) [27]	Childhood health conditions or states	Inception to December 2015	(i) Fixed-effect meta-analysis models (ii) Mixed- effects meta-	(i) Fixed-effect models: 1,073 mean utilities from 272 studies (ii) HLMs: (a) 279 mean utilities from 89	(i) Fixed-effect models: Mean utility from all direct and indirect valuation methods used in childhood population ^h weighted by	(i) Fixed-effect models: Pooling by ICD-10 delineated health condition categories and valuation method (ii) HLMs: ICD-10 delineated health condition

regression models (HLM)	studies; (b) 211 mean utilities from 67 studies	inverse of sample variance (ii) HLMs: Mean utility from (a) HUI3 and (b) VAS weighted by inverse of standard error. Studies weighted by number of respondents	categories; Valuation of hypothetical scenarios; Respondent type; Administration mode; Minimum age of sample; Country (developed vs. developing)
-------------------------	---	--	--

FE: Fixed-effect; RE: Random-effects; ME: Mixed-effects; HLM: Hierarchical linear model; SG: Standard gamble; CG: Chained gamble; TTO: Time trade-off; QWB: Quality of well-being scale; VAS: Visual analogue scale; RS: Rating scale; HALex: Health activities and limitations index; AQoL: Assessment of quality of life; ART: Anti-retroviral therapy

^a Moderate hepatitis C, compensated cirrhosis, decompensated cirrhosis, and post-liver transplant

^b Diabetic retinopathy, failed back surgery syndrome, post-herpetic neuralgia, phantom/residual limb pain, central neuropathy, and mixed neuropathy

^c Sample size depended on weights used and valuation methods included: 230 mean utilities from 29 studies when utilities weighted by sample size; 163 mean utilities from studies reporting standard deviation when utilities weighted by inverse of standard deviation; and 145 mean utilities when utilities weighted by sample size and VAS and TTO without full health top bound excluded.

^d Sample size depended on weights used and valuation methods included: 117 mean utilities from 20 studies when utilities weighted by sample size; 77 mean utilities from studies reporting standard deviation when utilities weighted by inverse of standard deviation; and 86 mean utilities when utilities weighted by sample size and VAS and TTO without full health top bound excluded.

^e Median utilities given minimal skew in distribution; median utilities weighted by pooled sample variance.

^f Samples within a study which were identical with respect to covariates included in regression but heterogeneous with respect to other variables were aggregated in baseline models but disaggregated in supplemental model.

^g 9 mean utilities from 3 studies using SG and 9 utilities from 3 studies using EQ-5D; 3 mean utilities per depression severity (mild, moderate, severe) for both SG and EQ-5D.

^h VAS, EQ-5D VAS, EQ-5D-Y VAS, TTO, SG, CG, QWB, 15D, 16D, 17D, EQ-5D, EQ-5D-Y, AQoL-5D, AQoL-6D, CHU9D, HUI2, HUI3, SF-6D, PAHOM and non-preference-based measures mapped to QALY

Table 3. Mixed-effect meta-regression by hierarchical linear model of utility values for paediatric cancer survivors measured by HUI2 and HUI3; 22 samples across 7 studies

Variable	Coefficient (SE)	Coefficient 95% CI	P-value
Constant (Utility for paediatric cancer survivors measured by HUI3 after 2000; self-response; self-administered questionnaire)	0.870 (0.043)	0.786 to 0.953	<0.001
Methodological factors			
Valuation method: Reference HUI3 vs. HUI2	0.043 (0.024)	-0.003 to 0.090	0.068
Respondent type: Reference self-response vs. proxy-response	0.029 (0.034)	-0.037 to 0.095	0.387
Administration mode: Reference self-administered vs. interview-administered	-0.042 (0.009)	-0.060 to -0.025	<0.001
Year of publication: Reference after 2000 vs. before or in 2000	-0.115 (0.040)	-0.194 to -0.036	0.004
Average predicted utility from model, weighted by inverse prediction standard error	0.937 (0.016)	0.905 to 0.969	

Figure 1. Fixed-effect meta-analysis of utility values for childhood cancer survivors assessed using Health Utilities Index 2 (HUI2) and Health Utilities Index 3 (HUI3)

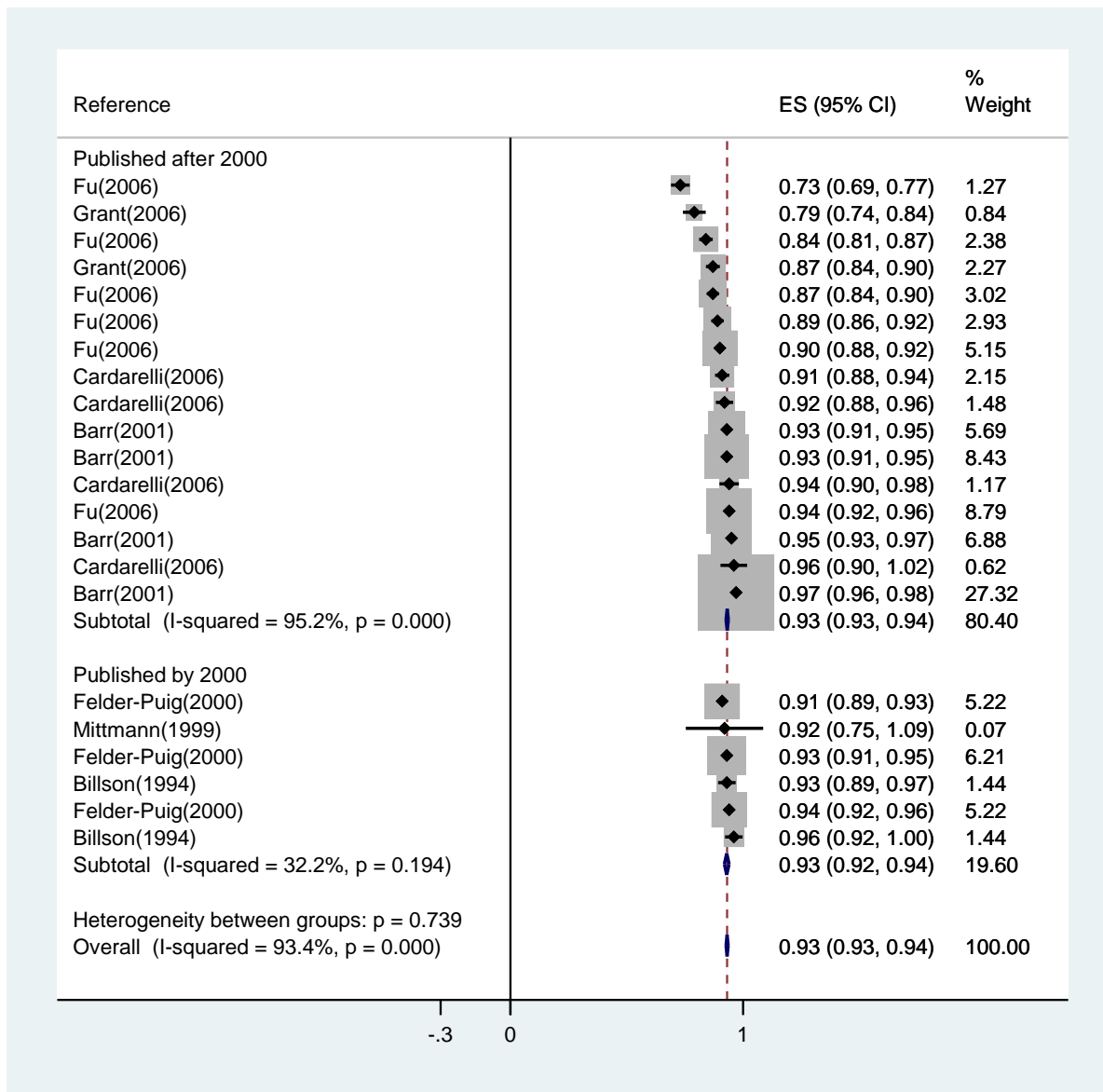
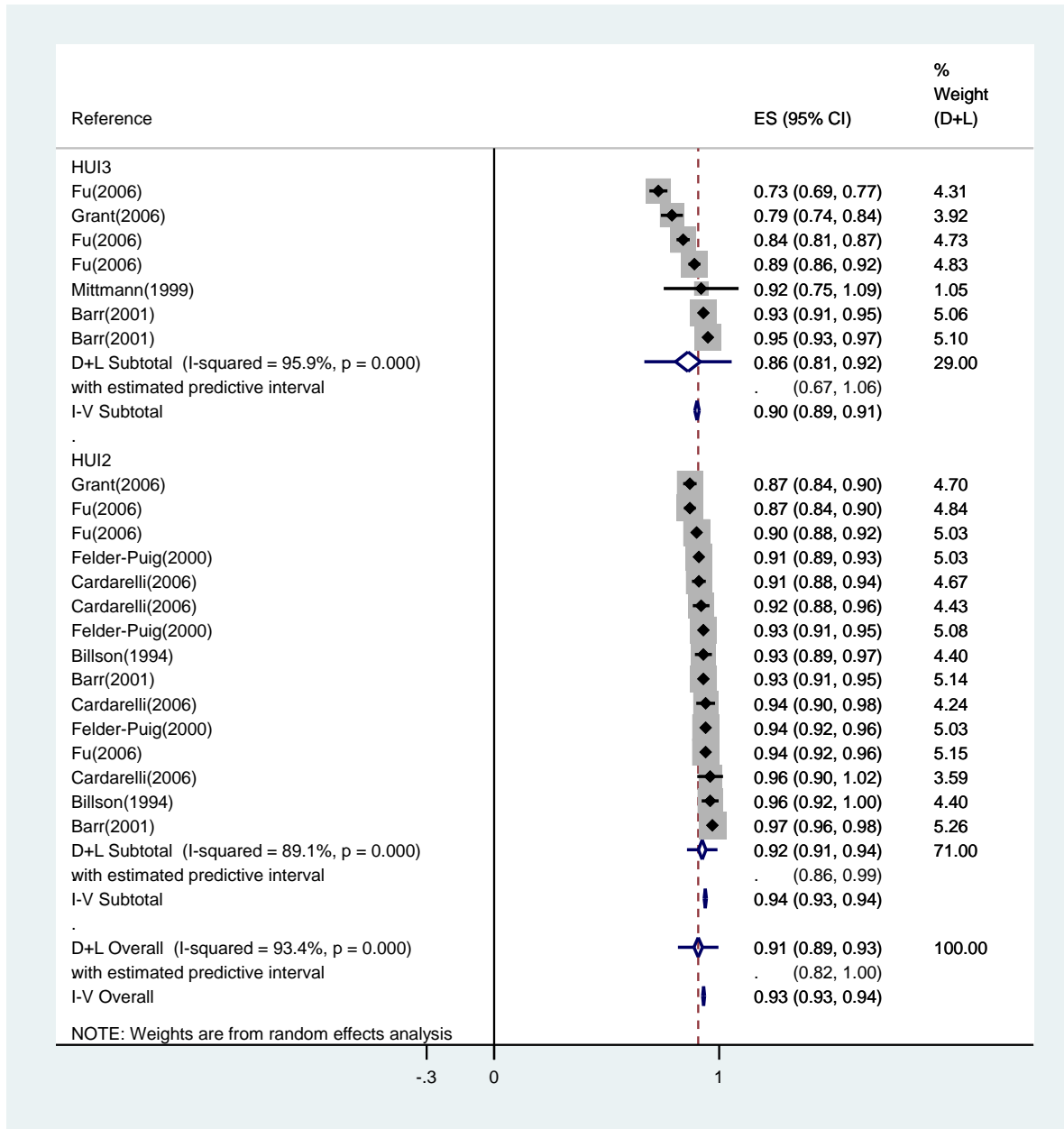


Figure 2. Random-effects and fixed-effect meta-analysis for childhood cancer survivors assessed using HUI2 and HUI3



Appendix A: Search strategy adopted by systematic review of childhood health state utility values

Database: PubMed, Embase, Web of Science, PsycINFO, EconLit, CINAHL, Cochrane Library		
Date: 26.01.2016		
Include articles with publication date up to 31.12.2015		
Limit search to title and abstract		
Search Category	Search Terms	
Utility Terms	1 Utilit* or disutilit* or HSUV	
	2 “quality adjusted life year*” or QALY or “quality-adjusted life year*” or “quality-adjusted life-year*”	
	3 OR (1 to 2)	
Indirect Valuation Method Terms	4 EQ-5D or “EQ 5D” or EQ5D or Euroqol or “Euro qol” or EQ-5D-Y or "EQ 5D Y"	
	5 Short-form survey-6D or short form 6D or SF-6D or “SF 6D” or SF6D	
	6 “health utilities index” or HUI	
	7 “quality of well being” or “quality of well-being” or QWB	
	8 16D Health-Related Quality of Life or 16D HRQoL or 17D Health-Related Quality of Life or 17D HRQoL	
	9 AQoL-6D or Assessment of Quality of Life-6D	
	10 “Child Health Utility 9 Dimension” or CHU9D or CHU-9D or “CHU 9D”	
	11 Adolescent Health Utility Measure or AHUM	
	12 15-dimensional instrument or 15 dimensional instrument or 15D	
	13 preference-based measure of HRQoL or preference based measure of HRQoL	
	14 multi-attribute utility instrument or multiattribute utility instrument	
	15 OR (4 to 14)	
	Direct Valuation Method Terms	16 Standard Gamble or standard-gamble
		17 Time trade off or time trade-off
		19 best worst scaling or best-worst scaling
19 Discrete choice experiment or discrete-choice experiment		
20 person trade off or person trade-off		
21 scoring algorithm or scoring-algorithm		
22 utility elicitation or direct elicitation		
23 OR (16 to 22)		
24 3 OR 15 OR 23		
Childhood Terms	25 Child* or adolesc* or kid or kids or youngster* or teen* or youth* or infant* or newborn* or perinat* or neonat* or “parent proxy”	
	26 Pediatri* or paediatri*	
	27 OR (25 to 26)	
Main Search	28 24 AND 27	
	29 Remove non-English Title and/or Abstract	
	30 Remove Duplicates Across Databases	

Source: Kwon J, Kim SW, Ungar WJ, Tsiplova K, Madan J, Petrou S. A Systematic Review and Meta-Analysis of Childhood Health Utilities. Med Decis Making. 2017;272989X17732990.[27]

Appendix B: Example data extraction sheet developed for systematic review of childhood health state utility values

Variables	Examples	Response
Reference		
Country/ jurisdiction/context	Country, sociodemographic context, clinic type and location, etc.	
Health condition/state or intervention state	Health state (specify duration), condition sub- category, medical treatment or intervention, phase of care (e.g. survivors, undergoing treatment), etc.	
Respondent type	Children, patients, parents, caregivers, nurses, physicians, other proxies, general population, etc.	
Age of target childhood group	Age at diagnosis, age at study, mean, median, range, min, max	
Size of study population		
Direct valuation method applied (if applicable)	Time trade-off, standard gamble, person trade-off, discrete choice experiment, etc.	
Indirect valuation method applied (if applicable)	16D, 17D, AQoL-6D, HUI2, HUI3, EQ-5D, EQ- 5D-Y, QWB, etc.	
Tariff if indirect valuation method applied	UK MVH, etc.	

Utility values	Mean, median, SD, IQ range, min, max, utility of control or reference group, utility difference between groups or after intervention, etc.			
Study design	Cross-sectional study, clinical trial, longitudinal prospective, internet survey, etc.			
Response quality	Response rate, information on dropouts, reasons for loss to follow-up, etc.			
Statistical method	Mean values, Linear multivariate regression model, etc.			
Study caveats and conclusion	Condition associated with lower utility; concern over sample size; concern over construct validity of instrument; etc.			
Quality appraisal:	Study design:	Definition of study group:	Representative study group:	Other issues:
Comment:				

Source: Kwon J, Kim SW, Ungar WJ, Tsiplova K, Madan J, Petrou S. A Systematic Review and Meta-Analysis of Childhood Health Utilities. *Med Decis Making*. 2017;272989X17732990.[27]