ScienceDirect

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

Methodology

Incorporating Prior Beliefs Into Meta-Analyses of Health-State Utility Values Using the Bayesian Power Prior

Anthony J. Hatswell, PhD

ABSTRACT

Objectives: Health-state utility values (HSUVs) directly affect estimates of Quality-Adjusted Life-Years and thus the cost-utility estimates. In practice a single preferred value (SPV) is often selected for HSUVs, despite meta-analysis being an option when multiple (credible) HSUVs are available. Nevertheless, the SPV approach is often reasonable because meta-analysis implicitly considers all HSUVs as equally relevant. This article presents a method for the incorporation of weights to HSUV synthesis, allowing more relevant studies to have greater influence.

Methods: Using 4 case studies in lung cancer, hemodialysis, compensated liver cirrhosis, and diabetic retinopathy blindness, a Bayesian Power Prior (BPP) approach is used to incorporate beliefs on study applicability, reflecting the authors' perceived suitability for UK decision making. Older studies, non-UK value sets, and vignette studies are thus downweighted (but not disregarded). BPP HSUV estimates were compared with a SPV, random effects meta-analysis, and fixed effects meta-analysis. Sensitivity analyses were conducted iteratively updating the case studies, using alternative weighting methods, and simulated data.

Results: Across all case studies, SPVs did not accord with meta-analyzed values, and fixed effects meta-analysis produced unrealistically narrow CIs. Point estimates from random effects meta-analysis and BPP models were similar in the final models, although BPP reflected additional uncertainty as wider credible intervals, particularly when fewer studies were available. Differences in point estimates were seen in iterative updating, weighting approaches, and simulated data.

Conclusions: The concept of the BPP can be adapted for synthesizing HSUVs, incorporating expert opinion on relevance. Because of the downweighting of studies, the BPP reflected structural uncertainty as wider credible intervals, with all forms of synthesis showing meaningful differences compared with SPVs. These differences would have implications for both cost-utility point estimates and probabilistic analyses.

Keywords: Bayesian Power Prior, EQ-5D, utility synthesis.

VALUE HEALTH. 2023; 26(9):1389–1397

Introduction

Health-state utility values (HSUVs) are crucial inputs to economic models, directly affecting the calculation of quality-adjusted life-years, and, consequently, the cost-utility of interventions. Recent literature on collection and use of HSUVs recommends a combination of systematic review and meta-analysis in the case of multiple potential (credible) values,[1-3] with information-sharing methods for clinical data that are also well established in health technology assessment (HTA).[4]

One recent example of the tension between different approaches to HSUV evidence is the National Institute for Health and Care Excellence (NICE) HTA of tepotinib for the treatment of non-small-cell lung cancer (NSCLC).[5] The pivotal VISION study for tepotinib included the EQ-5D-5L captured from 290 patients[6]; however, other sources were identified in the manufacturer submission, including clinical trials of other interventions in similar settings (n = 1034, 252, and 582), routine treatment data from 319 patients, and a widely used vignette study. When evaluating the cost-effectiveness of tepotinib, the directly measured EQ-5D clearly had the most relevance and was used exclusively as a Single Preferred Value (SPV), implicitly omitting the data from the other sources. Where some form of synthesis applied, the HSUVs with lesser (but nonzero) relevance would also have informed the parameters used in modeling.

When synthesis of HSUVs is appropriate, random effects meta-analysis (REMA), fixed effects meta-analysis (FEMA), and potentially, other methods, such as meta-regression, have been recommended.[3] Early HSUV syntheses tended to use REMA,[7-11] which appears to have subsequently become somewhat of a de facto standard in HSUV meta-analyses.[12-14] Exceptions to the use of REMA do exist—for instance, there are examples of HSUV meta-regression in both frequentist[15] and Bayesian[16] frameworks. The drawback of meta-regression being that it requires a large number of studies[17]; thus in many cases, it is not feasible given the limited number of HSUV observations available.

When synthesizing values, a limitation of REMA, FEMA, and some meta-regressions is that the relevancy of a study is not accounted for. Within these methods, studies are typically included or excluded depending on perceived relevancy—the excluded studies implicitly receive a weighting of 0. For included studies, the implication of ignoring relevancy is that, with REMA and FEMA, studies with a higher precision (due to either homogeneity in estimates, or larger sample sizes) receive higher weightings, irrespective of their relevance to the decision problem. Consequently, *ceteris paribus*, a larger study with less relevancy (eg, an older study focused on a different intervention) would have more influence on the results than a smaller, contemporary study with more relevancy. In the tepotinib example, the VISION study represents around 12% of the total number of patients, and would thus receive a low weighting using a REMA or FEMA.

In this article, the Bayesian Power Prior (BPP) is introduced as a method to differentially weight HSUVs during synthesis, providing a flexible approach to the inclusion of relevant evidence. These weights (derived from expert opinion of perceived relevancy) can then be formally incorporated in the decision-making process. In the article the example used is of HSUVs for a hypothetical NICE appraisal, that is, contemporary values using the relevant intervention, measured using EQ-5D, scored using a UK value set.[18] The more values depart from this ideal, the lower their relevance and thus the lower the weight given to the study. Sensitivity analyses are then conducted to illustrate differences in outcomes between methods, as well as the impact of different weighting schemes on the synthesized value.

Methods

Bayesian Power Prior

The BPP was introduced by Ibrahim and Chen.[19] The BPP in this context allows the synthesis of data sources with downweighting, enabling consideration of the degree of relevancy. The first example of BPP use was in water quality testing, in which relevancy was determined by data recency (less recent auxiliary data received a lower weighting).[20,21] Subsequently, BPPs have been used in synthesis and analysis of clinical trial data.[22-24] The use of the BPP, with relevancy again defined by recency, increased the statistical power and reduced the required sample size for prospective studies.

Mathematically, by using auxiliary data D_0, the conditional power prior of the current study θ is estimated by the historical likelihood $L(\theta|D_0)$ raised to the power a_0, and multiplied by the initial prior $\pi_0(\theta)$

$$\pi(\theta|D_0, a_0) \propto L(\theta|D_0)^{a_0}\pi_0(\theta)$$

In which π is the posterior distribution, and a_0 controls the degree of borrowing from D_0. Given primary data D, the conditional posterior distribution then becomes

$$\pi(\theta|D, D_0, a_0) \propto L(\theta|D)L(\theta|D_0)^{a_0}\pi_0(\theta)$$

When $a_0 = 0$, the auxiliary data have no influence on the posterior, whereas when $a_0 = 1$ the auxiliary data are not downweighted and information is fully borrowed. A universally accepted means of deriving or eliciting a_0 for each data source is yet to be determined in the literature. Approaches proposed in the literature include a_0 derived through comparing the similarity of data sets, known as commensurate power priors.[25] Alternatively, as a_0 is intuitive and does not require any Bayesian knowledge to elicit or interpret, expert beliefs on the relevancy of each study can be a useful means of derivation. Finally, there is potential for more derivation functions, for example, thresholds of similarity between studies, below which the data are discarded, also known as "test then pool."[23,26]

In this study the BPP is implemented using the normalized power prior,[26] multiplying the weights listed in Table 1 by the log of the probability distribution function (ie, the distribution from the data). This is implemented in the statistical package R, using the "rstan" package, which implements Bayesian analysis through the statistical software Stan.[27] The code used for the BPP analysis is presented in Box 1, which iterates through $1, \ldots, J$ observations (the number of studies) to estimate the BPP HSUV, incorporating the weights assigned. A half-normal prior is used for the between study heterogeneity with location parameter 0 and scale parameter 0.5, based on published simulation studies.[28]

Methods for Comparison, and Outcomes of Interest

The SPV and the literature-recommended methods[3] discussed in the introduction (REMA and FEMA) are used as benchmarks for comparison with the BPP approach. SPVs selected are those judged to best match the decision problem for each case study. REMA and FEMA are computed via the widely used "*metaphor*"[29] package within R.[30] Fixed effects models consider only the inverse variance of the standard error to weight estimates—in this context they assume a "true" utility and assume all variability is because of sampling error of the true utility. Random-effects models allow for between study variance in estimates and thus assume each study samples from the "true" utility distribution.

In all case studies, the estimands of interest are the population-level HSUV and associated uncertainty distribution; suitable for use in a cost-utility analysis.

Values for Synthesis, Perspective, and Derivation of Expert Beliefs

Four distinct examples are provided to explore different types of HSUVs to be synthesized. These take the perspective of the NICE methods guide,[18] which states a preference for EQ-5D-derived HSUVs based on a UK value set. All values used in the examples (HSUVs, measures of uncertainty, and weights) are listed in Table 1. Citations are given to the systematic reviews which we treat as the primary source (i.e. the searches are not performed, and data re-extracted). As results of Sampson et al. are not yet in press, citations are provided to the identified studies.[42-46] Also given in the table are justifications for power prior weights applied, which in this case are based on the authors beliefs as a proof of concept. These weights offer a starting position, which can be critiqued, adjusted according to the preferences of stakeholders, and ultimately values can be selected to provide HSUVs for decision making.

Case Studies

The first case study is in previously treated NSCLC, referenced in the introduction. As well as the tepotinib VISION study, 4 published estimates were identified in similar patient populations, although with imperfect relevancy and different interventions. Factors affecting the relevancy of these 4 studies beyond the intervention included differences in baseline age and disease subtypes between studies. The final estimate is a vignette study in which clinicians were interviewed in a structured format to derive their estimate for the HSUV.[31] Therefore, the weights given were 1 for the tepotinib study, lower for other clinical studies, and very low for the vignette study (Table 1).

Table 1. HSUVs used in meta-analysis and resulting meta-analyzed values for 4 case studies.

Study, sorted by year of publication	HSUV (SE)	Sample size	Power prior weight	Justification of weight used
Previously treated Non-Small Lung Cancer; values taken from NICE[5]				
Nafees et al[31]	0.674 (0.002)	10	0.01	Vignette study of 10 healthcare practitioners, approximately 15 years ago
Chouaid (2013)	0.74 (0.010)	319	0.4	Patients were treated in a non-trial setting, with a variety of therapies which predate current standard of care
NICE TA428: Pembrolizumab (2017)	0.74 (0.014)	1034	0.7	Although treated in a recent study, patients received immunotherapy, not targeted therapy
NICE TA484: Nivolumab, squamous (2017)	0.74 (0.010)	252	0.7	Although treated in a recent study, patients received immunotherapy, not targeted therapy
NICE TA655: Nivolumab, non-squamous (2020)	0.75 (0.002)	582	0.7	Although treated in a recent study, patients received immunotherapy, not targeted therapy
NICE TA789: Teptotinib VISION study (2022)	0.754 (0.016)	290	1	The contemporary study of patients treated with tepotinib
Hemodialysis; values taken from Cooper et al[32]				
Manns (2002)	0.6 (0.004)	128	0.3	Cross-sectional study of dialysis adequacy and quality of life in Canadian patients, approximately 20 years old
Gorodetskaya (2005)	0.54 (0.019)	271	0.3	US study estimating quality of life in dialysis patients with multiple metrics used
Lee (2005)	0.44 (0.032)	99	0.8	Well conducted (although dated) Welsh study
Manns (2009)	0.69 (0.009)	51	0.4	Canadian study, investigating if nocturnal dialysis improves quality of life
Briggs (2016)	0.75 (0.006)	1767	0.9	Utility analysis of a large multinational clinical study
Jardine (2017)	0.78 (0.017)	200	0.5	Clinical to improve quality of life of patients by altering dialysis duration, conducted in Oceania, China, and Canada
Pan (2018)	0.75 (0.006)	315	0.2	Single-center cross-sectional study in China to investigate factors related to better quality of life
Wong (2019)	0.73 (0.009)	135	0.2	Chinese study, reporting the utility of different groups undergoing dialysis
Wong (2019)	0.78 (0.014)	41	0.2	
Wong (2019)	0.79 (0.010)	118	0.2	
Compensated cirrhosis in Hepatitis C; values taken from Saeed et al[33]				
Siebert (2001)	0.74 (0.026)	74	0.6	Data approximately 20 years old; however, collected in a European country (Germany) using EQ-5D
Chong (2003)	0.74 (0.041)	24	0.5	Data are nearly 20 years old, and collected/scored using the Canadian EQ-5D-3L value set
Bjornsson (2009)	0.749 (0.024)	76	0.7	Data are over 10 years old, and collected/scored using the Swedish EQ-5D-3L value set
Wright (2009)	0.550 (0.054)	40	0.9	UK data; however, it is over 15 years old, and using EQ-5D-3L
Kieran (2012)	0.600 (0.043)	68	0.8	Only an abstract; however, it uses EQ-5D-3L in a similar healthcare system and population
Vellopoulou (2014)	0.730 (0.047)	23	0.9	The Netherlands data, and relatively recent, using EQ-5D-3L
Pol (2015)	0.67 (0.30)	101	0.7	Multi country EQ-5D-3L, unclear which value set was used
Vargas (2015)	0.682 (0.285)	9	0.6	Collected in Chile, only has 9 patients; therefore, already a very uncertain estimate
Kaishima (2016)	0.774 (0.028)	67	0.5	Although recent, data were collected in Japan, and scored using the Japanese EQ-5D-3L tariff
Blindness due to diabetic retinopathy, values identified by Sampson et al[14]				
Coffey (2002)	0.534 (0.038*)	62	0.3	Alternative models fit using the Self-Administered Quality of Well Being index to estimate utilities in a diabetic population in the USA
Coffey (2002)	0.347 (0.032*)	90	0.4	
Coffey (2002)	0.510 (0.026*)	129	0.4	
Coffey (2002)	0.361 (0.026*)	135	0.4	
Ohsawa (2003)	0.76 (0.036)	60	0.5	Standard gamble (higher) and SF-36 (lower) based estimates of quality of life from hospitalized Japanese diabetic patients
Ohsawa (2003)	0.2 (0.036)	60	0.2	
Huang (2007)	0.39 (0.003)	519	0.5	Time trade-off study conducted in the USA, in diabetic patients
Huang (2007)	0.38 (0.013)	701	0.4	
Chin (2008)	0.39 (0.015)	473	0.5	Time trade-off utility values in an elderly population in the USA
Lee (2008)	0.4 (0.076)	25	0.5	Standard gamble valuations of blindness, each taken from different populations with a variety of visual impairments
Lee (2008)	0.71 (0.066)	25	0.3	
Lee (2008)	0.43 (0.057)	33	0.5	
Lee (2008)	0.74 (0.045)	33	0.3	
Sullivan (2016)	0.613 (0.017)	593	0.7	EQ-5D utilities using the UK tariff in a USA community data set of patients with diabetes related comorbidities

*SD not reported; therefore, assumed to be the mean of other SDs in the disease area.

BOX 1. Stan Code for the Implementation of the Bayesian Power Prior.

```
data {
  int<lower=0> J;                 // Number of studies
  real<lower=0, upper=1> y[J];    // Estimated HSUV
  real<lower=0> sigma[J];         // Uncertainty around HSUV
  real<lower=0> weight[J];        // BPP weight
}

parameters {
  real<lower=0> mu;
  real<lower=0> theta[J];
  real<lower=0> tau;
}

model {
  mu ~ normal(0, 5);
  tau ~ normal(0, 0.5);
  for (n in 1:J) {
    theta[n] ~ normal(mu, tau);
    target += weight[n] * normal_lpdf(y[n] | theta[n], sigma[n]);
  }
}
```

The second case study uses the results from a published systematic review of HSUVs in chronic kidney disease.[32] The HSUV of interest was that of patients receiving hemodialysis. In this case, 10 available HSUVs (published 1992-2019) used a variety of quality-of-life instruments, which affects the suitability for meta-analysis. All values of a_0 are <1 in this case, because there is no directly relevant study that is specific to the patient population, with assigned weights and rationales given in Table 1.

The third case study uses results from a published systematic review and meta-analysis of HSUVs in Hepatitis C, specifically looking at the health state of compensated liver cirrhosis.[33] None of the 9 values identified in the review are from a contemporary study of an intervention; studies were conducted in different countries, and all studies were scored using different EQ-5D value sets, reducing their relevancy. Consequently, all assigned weights of a_0 are again <1 because of the limited applicability to a UK HTA.

Finally, the fourth case study uses results from an ongoing systematic review of utilities in diabetic retinopathy.[14] The HSUV of interest was that of blindness in patients with diabetic retinopathy (DR blindness). Fourteen published estimates are available from a variety of settings and instruments. Distinct from the other case studies, weights were given using the expert opinion of the systematic review's first author, rather than the author of this article.

Sensitivity Analysis

For all case studies, results are presented including iterative updating of estimates each year additional HSUVs are published—beginning when at least 2 HSUVs are available for synthesis, and ending with the final model using all data. This sensitivity analysis illustrates how the methods vary with differing numbers of available HSUVs.

Following the case studies, as a demonstration of the impact a_0 can have on BPP results, a series of arbitrary weights are applied in each case study. These weights are based on either the mean or the standard error of the estimated HSUV, to show the range of outcomes that would be achieved under different weighting schemes. The scoring systems used in the sensitivity analysis are entirely illustrative, and are not based on beliefs regarding the suitability of data, simply on the ranked raw values from the studies. Sensitivities conducted include ranking values largest to smallest (and vice versa) for both the mean value and standard error, and applying weights to the studies based on these ranks. Although arbitrary, this sensitivity analysis serves to illustrate the impact alternative weights have on the synthesized estimates.

An illustration of differences between the methods is then presented using simulated data for 3 examples that synthesize 5 studies. In example A, the studies are equally sized, with the first study having HSUV of 0.8 and BPP weight of 1. HSUVs then reduce by 0.05, and BPP weight by 0.2 for each successive study. Example B extends this example by replicating the characteristics of study 1 for studies 2 to 4, leaving study 5 unchanged as a divergent estimate with a low BPP weight. Example C again uses the same setup as example A, but with study precision (the combination of standard error, and patient numbers) also decreasing from study 1 to study 5.

Results

Case Studies

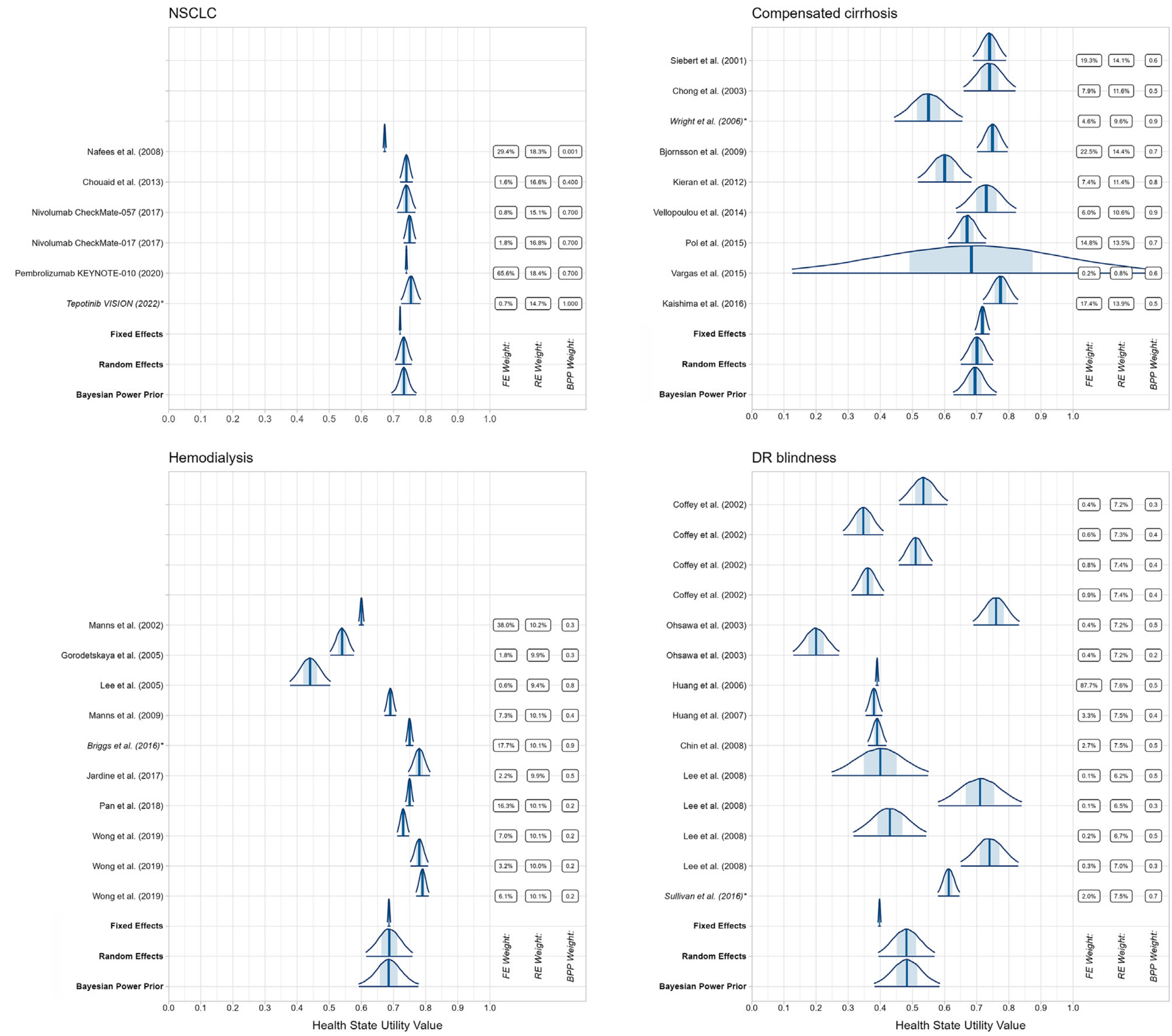
Figure 1 shows the inputs and results of the case studies, using an "MCMC area plot" (from the *"bayesplot"* R package). In each case, the calculated weight of each study is given for the REMA and FEMA studies, the weight used in BPP estimates, the mean, and 95% interval for all studies and synthesis methods (shown as the 95% CI for frequentist estimates, and 95% credible interval for Bayesian estimates). Although there are similarities between methods, there are also important subtleties, which led to differences in point estimates and uncertainty estimates across the different case studies.

The HSUV derived from the tepotinib VISION study is selected for the SPV because of its obvious relevancy. This SPV is the highest HSUV estimate from the 6 available (0.754 [CI 0.724-0.784]). REMA and FEMA provide similar results (estimated HSUVs of 0.731 [CI 0.706-0.757] and 0.719 [CI 0.716-0.721]; Table 1). FEMA does not allow for study effects leading to a high weighting of KEYNOTE-010 and Nafees et al (>95% of the weight in combination), leading to lower estimated uncertainty than REMA. The BPP approach leads to an estimated HSUV of 0.732 (credible interval 0.694-0.771); closer to the REMA results, with a slightly wider uncertainty bound.

In the hemodialysis case study (Fig. 1 and Table 1), the SPV HSUV is Briggs et al[34] (0.75 [0.738-0.762]) as a published analysis of patient-reported outcomes using a large data set. Despite this study being preferred, all included studies have issues with relevancy. Some studies include European experience and are given higher weights, whereas other studies represent Chinese and US outpatient settings, receiving lower weights. REMA and FEMA point estimates were similar, but with FEMA estimating lower uncertainty (0.687, [0.615-0.759]; 0.686, [0.681-0.691]). The BPP approach leads to an estimated HSUV of 0.685 (0.592-777). The higher uncertainty in REMA and BPP is driven by the divergent estimates, and using the BPP, the low weights assigned to the pool of studies as a whole decreased confidence in the finding.

In the compensated cirrhosis case, no study has a particularly large sample size or clear applicability to a contemporary HTA. The SPV HSUV is Wright et al[35] (0.550 [0.445-0.655]) in being based on a UK study, although given the study age, it may not represent contemporary patient experiences. REMA and FEMA estimates are 0.701 [0.651-0.751] and 0.718 [0.695-0.740]. The BPP approach leads to an estimated HSUV of 0.695 [0.628-0.761]. All meta-analyzed HSUV point estimates are similar (although with differences in uncertainty intervals).

Figure 1. Inputs and results of different approaches to the synthesis of health-state utility values for 4 different conditions.



BPP indicates Bayesian Power Prior; FE, fixed effect; NSCLC, non-small-cell lung cancer; RE, random effect.
*Single Preferred Value shown in italics and marked with asterisk.

The DR blindness case study is complex with 14 estimates from 7 studies that have considerable differences between study populations, settings, and instruments used. The SPV HSUV is Sullivan and Ghushchyan[36] (0.613 [0.580-0.646]) because a UK tariff is used and the sample size is relatively large, leading also to the largest BPP weight in this case study. REMA and FEMA estimates are 0.481 (0.394-0.569) and 0.398 (0.393-0.402), whereas the BPP approach leads to an estimated HSUV of 0.483 (0.382-0.584). BPP weights assigned in this case study had the lowest mean of the 4 case studies (0.42), implying the lowest relevancy and therefore highest uncertainty overall. This uncertainty is not reflected in the FEMA, with 88% of the weight derived from Huang et al,[37] which has a particularly small standard error. The BPP and REMA both show wider uncertainty bounds (with the BPP again widest) and a higher estimated HSUV than FEMA.

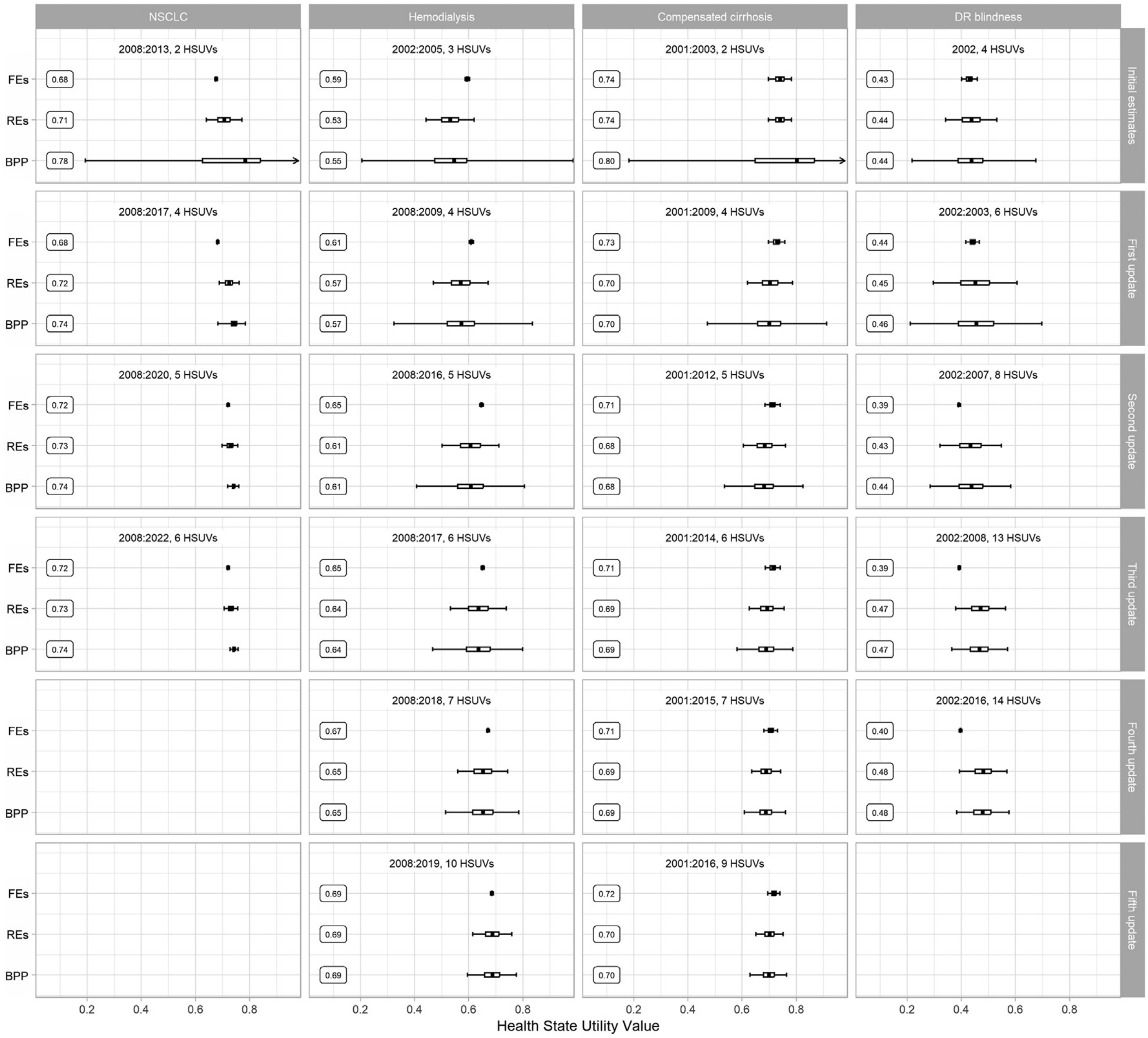
Sensitivity Analyses

The impact of the different methods is demonstrated in Figure 2 which shows the path to the final estimates, assuming iterative updates in any year in which an estimate was published (studies are ordered by publication year in Table 1). When evidence is sparse—with low numbers of HSUVs in the early estimates—each of the approaches have clear differences in point estimates and uncertainty intervals. As the number of HSUVs increases, estimates converge for REMA and the BPP, with FEMA being noticeably different. With the low weights assigned to the earliest HSUVs for several of the case studies, the credible intervals the BPP estimates often exceed 1, and only reduce with the introduction of higher quality studies, which have larger patient numbers (thus influence) and higher weights. This is in contrast with FEMA in which precise (although not necessarily reliable) estimates are available from a low number of studies, without reference to relevance.

Figure 3 demonstrates the impact different approaches to BPP weights can have on both estimates and uncertainty intervals, with notable differences between the stylized approaches presented. As would be expected, in cases which studies are relatively similar—for example NSCLC—the exact weights used makes relatively little difference because the precision of the studies carries substantial influence. Differences are more notable, however, in cases which the range is greater, such as in DR blindness; the difference in point estimate between highest and lowest approaches is 0.095. Impacts can also be seen even in the way in which values are assigned within a ranking—for example, if the same rankings are used, but the weights then squared. The width of credible intervals is also clearly affected by weights, although there remains a limit to how far these can be reduced, given downweighting of studies.

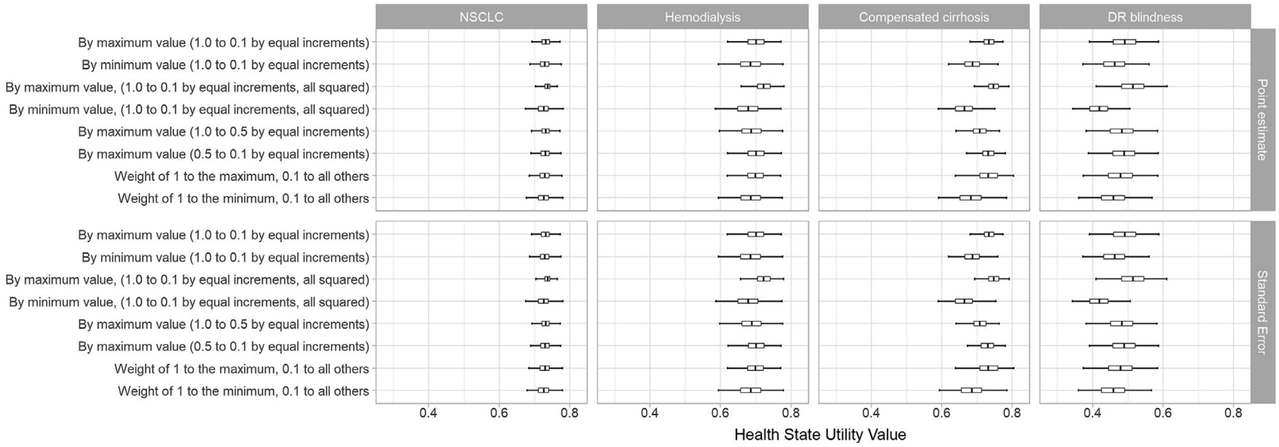
The results from simulated examples A, B, and C are shown in Figure 4, with divergences in REMA, FEMA, and BPP estimates—in these examples, due to HSUV and BPP weight being correlated. In example A, 5 otherwise identical studies (hence, equal weight in REMA/FEMA) aside from point estimates are presented, with the BPP weights shifting the distribution of the posterior. Example B shows a situation in which 1 study is downweighted greatly compared with the others. This difference leads to the same estimate for REMA/FEMA—although with more uncertainty in REMA regarding the true value. The BPP, however, gives higher estimates mean (due to the weights used) and has a skewed distribution—acknowledging the potential for the HSUV to be lower (because the study is not discarded). In example C, the differences in study precision leads to a divergence in estimates between REMA and FEMA, due to the differences in point estimates. The BPP model then reflects not only the heterogeneity between estimates, but also the impact of weights in again skewing toward the higher estimates of HSUV.

Figure 2. Comparison of methods for meta-analysis of health-state utility values; mean, interquartile range, and 95% interval, over yearly iterations.



BPP indicates Bayesian Power Prior; FE, fixed effect; RE, random effect.

Figure 3. Alternative weights based on mean or standard error used as sensitivity analyses in the Bayesian Power Prior.



Discussion

Case Studies

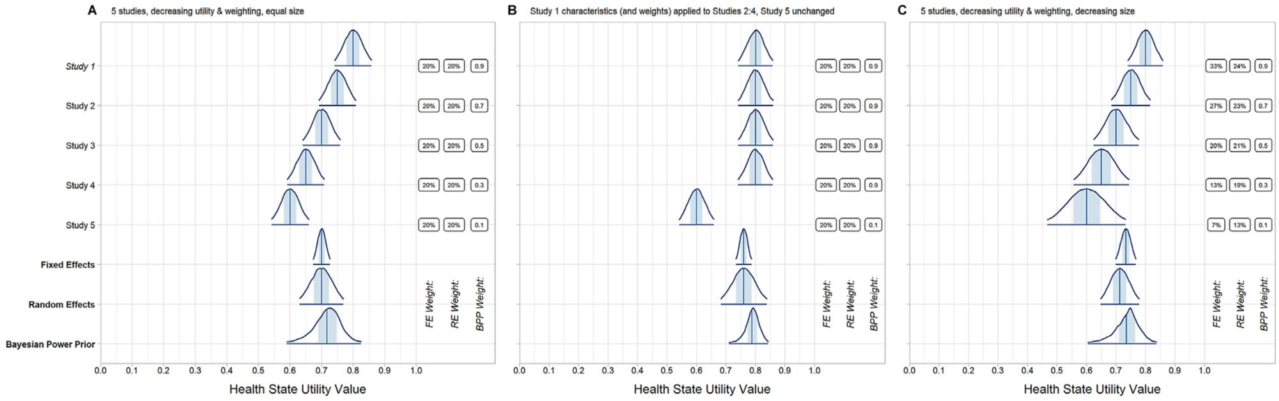
When comparing the results of all case studies (Fig. 1), the SPV HSUV often does not align with meta-analyzed results; in only 1 case study (NSCLC) is the SPV point estimate in the interquartile range of the synthesis methods. In 2 case studies (hemodialysis and DR blindness), SPV has the highest point estimate by a considerable margin, and in the final case study (compensated cirrhosis), it is the lowest. Within the synthesis methods, point estimates including all studies were generally similar between REMA, FEMA, and the BPP approach. Nevertheless, substantial differences were seen in the width of uncertainty estimates because all FEMA models suggested low uncertainty in all cases. This contrasts to both REMA and BPP approaches that estimated greater uncertainty, with the BPP having slightly wider uncertainty bounds—reflecting the downweighting provided by a_0.

In terms of individual methods, FEMA makes assumptions about studies that appear difficult to justify in utility analyses, given the results from the 4 case studies examined. Ultimately, clinical trials include patients with different characteristics, receiving different treatments, and in different settings/countries. These differences in study effects render it difficult to justify the assumption of a single "true" utility, because this would seemingly be affected by any number of (known) differences between studies. For this reason, REMA would seem as a more appropriate choice of frequentist approach should a meta-analysis of HSUVs be deemed appropriate.

Although the final case study models gave similar estimates for REMA and the BPP, clear differences between the methods emerged in the iterative updating (Fig. 2) and using simulated data (Fig. 4). These examples show cases in which BPP offers a theoretical advantage over REMA, by considering the relevancy of the information being entered into the estimator. By downweighting auxiliary data, the primary data can maintain prominence, while also allowing the inclusion of other sources. In a practical setting, REMA's equal treatment of all studies is problematic for meta-analyses of HSUVs—for example, the Nafees et al study in the NSCLC case study; interviews of 10 clinicians. Due to homogenous estimates being given, the standard error is extremely small, and as a result, in REMA, the study has similar influence to a clinical trial of 582 patients.

The final difference with the use of BPP is in the distribution of uncertainty, in which the BPP can result in asymmetric distributions. This is clearly seen in the simulated data (Fig. 4), in which the correlation between weight and utility led to skewed distributions in the posterior—particularly in examples A and B in which the use of BPP weights have an obvious impact on the model result. Across the simulated examples, the possibility of low

Figure 4. Simulated examples illustrating differences between meta-analysis and BPP approaches.



BPP indicates Bayesian Power Prior; FE, fixed effect; RE, random effect.

HSUVs is not discounted, but deemed to be less likely by virtue of the Bayesian approach—something not seen/possible with the meta-analysis approaches.

The main finding of the sensitivity analyses, varying the way studies were weighted, was that it is not only the relative ranking of studies that is important (which is self-evident), but that results are also affected by the absolute value of BPP weights. The gap between weights (demonstrated by squaring the resulting weights) influenced estimates, as did using weights of 0.1 to 0.5, versus 0.5 to 1. Although stylized approaches, this sensitivity analysis demonstrates that justification should not only be given for the ranking of BPP weights, but also the absolute values used.

Derivation of a_0, That Is, Setting of Priors

The main challenge for the BPP approach is assignment of weights. For the case studies, “expert opinion” was required to be used because of wide variation in the suitability of estimates. This however, was itself a judgment that precluded the use of commensurate power priors. Careful consideration of the appropriate implementation of the BPP is required, should its use be planned, including who will provide weights, and how.[38]

In cases which expert opinion is to be used, although frameworks have been developed for structured expert elicitation,[39] this would be an additional step that must be planned for and tailored to the context of deriving a_0. Further research is also required on the appropriate use of the method, because even the bounds and perspective of a_0 are open to interpretation; should the most applicable study be a weight of 1 and other values be relative to this? Or should a_0 be determined according to the fit to a desired objective—in the case studies presented, the NICE reference case? The sensitivity analyses conducted show that such questions do matter.

The Use of Meta-Analysis for HSUVs

Although SPV is commonly used in HTA, there are clear theoretical (and practical) advantages of meta-analysis over SPV, despite the differences in estimates produced by meta-analysis methods. These advantages are in keeping with the approaches identified on information sharing in HTA.[4] Importantly, meta-analysis methods also characterize the uncertainty between estimates of HSUVs, which may otherwise be omitted. As such, evidence synthesis methods, even if only used as used as sensitivity analyses, could aid decision making by reflecting both the between study heterogeneity and the structural uncertainty around estimates.

Should the BPP approach be used, further developments in the implementation are required for use in HTA. For example, the case studies presented involve only single health states rather than multiple states (such as predisease and postdisease progression in oncology), which are generally required in cost-utility analyses. Given that studies are likely to provide estimates from >1 health state (or as in the cirrhosis example, may provide multiple estimates), hierarchical models may be required to allow for study level correlation. There may also be a desire to make the models used fully Bayesian by including priors, for example, that HSUVs be bounded either by 1, or that of the general population. Similarly, strong priors could be included in hierarchical models, such as utility strictly decreasing upon disease progression.

Beyond the BPP, an additional difficulty is in identifying values for synthesis. In general, the reporting around HSUVs in the literature is poor, with often only mean values being given, and little reporting of dispersion. Better reporting of HSUVs with at least the SD, standard error, or 95% confidence interval will therefore be required in the literature for meta-analysis to become a standard. This is recognized as good practice, for example, as Item 18 in the CHEERS (Consolidated Health Economic Evaluation Reporting Standards) statement,[40] and Item 22 in the CHEERS 2022 statement.[41]

Conclusions

The use of the BPP approach is not a panacea for the problem of estimating HSUVs for use in HTA. In being able to explicitly downweight less relevant studies, however, the advantages of meta-analysis (the use of all applicable data and appropriate uncertainty estimates) are retained without these auxiliary sources overpowering the most applicable source(s). Although further development is required, the advantages of BPP appear to be clear and should facilitate more reflective estimates of HSUVs for use in economic modeling—and ultimately—decisions on whether to adopt novel technologies.

Article and Author Information

Accepted for Publication: April 28, 2023

Published Online: June 21, 2023

doi: https://doi.org/10.1016/j.jval.2023.04.012

Author Affiliations: Delta Hat Limited, Nottingham, England, UK (Hatswell); Department of Statistical Science, University College London, London, England, UK (Hatswell).

Correspondence: Anthony Hatswell, PhD, Delta Hat Limited, Bramley House, Bramley Road, Nottingham NG10 3SX, England, United Kingdom. Email: ahatswell@deltahat.com

Author Contributions: *Concept and design:* Hatswell
Acquisition of data: Hatswell
Analysis and interpretation of data: Hatswell
Drafting of the article: Hatswell
Critical revision of the article for important intellectual content: Hatswell
Provision of study materials or patients: Hatswell
Administrative, technical, or logistic support: Hatswell

Conflict of Interest Disclosures: The author reported no conflicts of interest.

Funding/Support: The author received no financial support for this research.

Acknowledgment: The author would like to thank the following: W Sullivan, A Bullement, G Baio, C Sampson, S Palmer, P Tappenden, H Wheat, D Lee, the audiences of “R for HTA,” HESG and ISPOR Vienna, S Lewis, D Burns, and M Sculpher. The final work is also the result of superb peer reviewers, who massively improved the illustration and communication of the approach.

REFERENCES

1. Papaioannou D, Brazier J, Paisley S. Systematic searching and selection of health state utility values from the literature. *Value Health*. 2013;16(4):686–695.
2. 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.
3. Petrou S, Kwon J, Madan J. A practical guide to conducting a systematic review and meta-analysis of health state utility values. *Pharmacoeconomics*. 2018;36(9):1043–1061.
4. Nikolaidis GF, Woods B, Palmer S, Soares MO. Classifying information-sharing methods. *BMC Med Res Methodol*. 2021;21(1):107.
5. Tepotinib for treating advanced non-small-cell lung cancer with MET gene alterations. National Institute for Health and Care Excellence. https://www.nice.org.uk/guidance/indevelopment/gid-ta10630. Accessed June 11, 2021.
6. Paik PK, Felip E, Veillon R, et al. Tepotinib in non–small-cell lung cancer with *MET* Exon 14 skipping mutations. *N Engl J Med*. 2020;383(10):931–943.
7. Tengs TO, Lin TH. A meta-analysis of utility estimates for HIV/AIDS. *Med Decis Making*. 2002;22(6):475–481.

8. Tengs TO, Lin TH. A meta-analysis of quality-of-life estimates for stroke. *Pharmacoeconomics*. 2003;21(3):191–200.
9. Liem YS, Bosch JL, Arends LR, Heijenbrok-Kal MH, Hunink MG. Quality of life assessed with the Medical Outcomes Study Short Form 36-Item Health Survey of patients on renal replacement therapy: a systematic review and meta-analysis. *Value Health*. 2007;10(5):390–397.
10. 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):733–741.
11. 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):853–868.
12. Aceituno D, Pennington M, Iruretagoyena B, Prina AM, McCrone P. Health state utility values in schizophrenia: a systematic review and meta-analysis. *Value Health*. 2020;23(9):1256–1267.
13. Blom EF, Haaf KT, de Koning HJ. Systematic review and meta-analysis of community- and choice-based health state utility values for lung cancer. *Pharmacoeconomics*. 2020;38(11):1187–1200.
14. 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.
15. Peasgood T, Ward SE, Brazier J. Health-state utility values in breast cancer. *Expert Rev Pharmacoecon Outcomes Res*. 2010;10(5):553–566.
16. Hatswell AJ, Burns D, Baio G, Wadelin F. Frequentist and Bayesian meta-regression of health state utilities for multiple myeloma incorporating systematic review and analysis of individual patient data. *Health Econ*. 2019;28(5):653–665.
17. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, United Kingdom: John Wiley & Sons, Ltd. http://onlinelibrary.wiley.com/doi/10.1002/9780470712184.fmatter/summary. Accessed June 11, 2022.
18. NICE health technology evaluations: the manual. National Institute for Health and Care Excellence. https://www.nice.org.uk/process/pmg36/resources/nice-health-technology-evaluations-the-manual-pdf-72286779244741. Accessed June 11, 2022.
19. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Stat Sci*. 2000;15(1):46–60.
20. Duan Y, Smith EP, Ye K. Using power priors to improve the binomial test of water quality. *J Agric Biol Environ Stat*. 2006;11(2):151–168.
21. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. *Stat Med*. 2015;34(28):3724–3749.
22. Banbeta A, van Rosmalen J, Dejardin D, Lesaffre E. Modified power prior with multiple historical trials for binary endpoints. *Stat Med*. 2019;38(7):1147–1169.
23. Dejardin D, van Rosmalen J, Lesaffre E. Including Historical Data in the Analysis of Clinical Trials Using the Modified Power Prior Practical Considerations for Survival Models. *Stat Methods Med Res*. 2018;27(10):3167–3182.
24. Pan H, Yuan Y, Xia J. A calibrated power prior approach to Borrow information from historical data with application to biosimilar clinical trials. *J R Stat Soc C Appl Stat*. 2017;66(5):979–996.
25. Hong H, Fu H, Carlin BP. Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta-analysis. *J R Stat Soc C Appl Stat*. 2018;67(4):1047–1069.
26. Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. *Stat Med*. 2009;28(28):3562–3566.
27. RStan: the R Interface to Stan. R Package Version 2.14.1. Stan Development Team. http://mc-stan.org/. Accessed March 9, 2017.
28. Röver C, Bender R, Dias S, et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Res Synth Methods*. 2021;12(4):448–474.
29. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1–48.
30. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org. Accessed June 11, 2022.
31. Nafees B, Stafford M, Gavriel S, Bhalla S, Watkins J. Health state utilities for non small cell lung cancer. *Health Qual Life Outcomes*. 2008;6:84.
32. Cooper JT, Lloyd A, Sanchez JJG, Sörstadius E, Briggs A, McFarlane P. Health related quality of life utility weights for economic evaluation through different stages of chronic kidney disease: a systematic literature review. *Health Qual Life Outcomes*. 2020;18(1):310.
33. Saeed YA, Phoon A, Bielecki JM, et al. A systematic review and meta-analysis of health utilities in patients with chronic hepatitis C [published correction appears in Value Health. 2020;23(7):977]. *Value Health*. 2020;23(1):127–137.
34. Briggs AH, Parfrey PS, Khan N, et al. Analyzing health-related quality of life in the EVOLVE trial: the joint impact of treatment and clinical events. *Med Decis Making*. 2016;36(8):965–972.
35. Wright M, Grieve R, Roberts J, Main J, Thomas HC. UK Mild Hepatitis C Trial Investigators. Health benefits of antiviral therapy for mild chronic hepatitis C: randomised controlled trial and economic evaluation. *Health Technol Assess Winch Engl*. 2006;10(21):1–113, iii.
36. Sullivan PW, Ghushchyan VH. EQ-5D scores for diabetes-related comorbidities. *Value Health*. 2016;19(8):1002–1008.
37. Huang ES, Shook M, Jin L, Chin MH, Meltzer DO. The impact of patient preferences on the cost-effectiveness of intensive glucose control in older patients with new-onset diabetes. *Diabetes Care*. 2006;29(2):259–264.
38. Bojke L, Soares M, Claxton K, et al. Developing a reference protocol for structured expert elicitation in health-care decision-making: a mixed-methods study. *Health Technol Assess*. 2021;25(37):1–124.
39. Gosling JP. SHELF: the Sheffield elicitation framework. In: Int S Oper Res Manag Sci, Dias LC, Morton A, Quigley J, eds. Elicitation: The Science and Art of Structuring Judgement. Berlin, Germany: Springer International Publishing. 2018:61-93.
40. Husereau D, Drummond M, Petrou S, 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):231–250.
41. Husereau D, Drummond M, Augustovski F, et al. Consolidated health economic evaluation reporting standards (CHEERS) 2022 explanation and elaboration: a report of the ISPOR CHEERS II good practices task force [published correction appears in Value Health. 2022;25(6):1060]. *Value Health*. 2022;25(1):10–31.
42. Coffey JT, Brandle M, Zhou H, et al. Valuing Health-Related Quality of Life in Diabetes. *Diabetes Care*. 2002;25(12):2238–2243. https://doi.org/10.2337/diacare.25.12.2238.
43. Ohsawa I, Ishida T, Oshida Y, Yamanouchi K, Sato Y. Subjective health values of individuals with diabetes in Japan: comparison of utility values with the SF-36 scores. *Diabetes Res Clin Pract*. 2003;62(1):9–16. https://doi.org/10.1016/s0168-8227(03)00145-1.
44. Huang ES, Brown SES, Ewigman BG, Foley EC, Meltzer DO. Patient perceptions of quality of life with diabetes-related complications and treatments. *Diabetes Care*. 2007;30(10):2478–2483. https://doi.org/10.2337/dc07-0499.
45. Lee BS, Kymes SM, Nease RF, Sumner W, Siegfried CJ, Gordon MO. The impact of anchor point on utilities for 5 common ophthalmic diseases. *Ophthalmology*. 2008;115(5):898–903.e4. https://doi.org/10.1016/j.ophtha.2007.06.008.
46. Chin MH, Drum ML, Jin L, Shook ME, Huang ES, Meltzer DO. Variation in Treatment Preferences and Care Goals Among Older Diabetes Patients and Their Physicians. *Med Care*. 2008;46(3):275–286. https://doi.org/10.1097/MLR.0b013e318158af40.