Article:
Ara, R.M. orcid.org/0000-0002-7920-1707, Brazier, J. and Young, T. (2017) Recommended Methods for the Collection of Health State Utility Value Evidence in Clinical Studies. PharmacoEconomics, 35 (Suppl 1). pp. 67-75. ISSN 1170-7690

https://doi.org/10.1007/s40273-017-0549-6

RECOMMENDED METHODS FOR THE COLLECTION OF HSUV EVIDENCE IN CLINICAL STUDIES

Running title: Collecting HSUVs in clinical studies

Authors: Roberta Ara MSc[1], John Brazier PhD[1], Tracey Young PhD[1].

[1]University of Sheffield

School of Health and Related Research (ScHARR)

Regent Court

Regent Street

Sheffield

UK

Corresponding author: John Brazier: j.e.brazier@sheffield.ac.uk

Abstract

A conceptual model framework and an initial literature review are invaluable when considering what health state utility values (HSUV) are required to populate health states in decision models. They are the recommended starting point early within a research and development programme, and before development of Phase 3 trial protocols.

While clinical trials can provide an opportunity to collect the required evidence, their appropriateness should be reviewed against the requirements of the model structure taking into account population characteristics, time horizon and frequency of clinical events. Alternative sources such as observational studies or registries may be more appropriate when evidence describing changes in HSUVs over time or rare clinical events is required. Phase 4 clinical studies may provide the opportunity to collect additional longitudinal real-world evidence. Aspects to consider when designing the collection of the evidence include patient and investigator burden, whom to ask, the representativeness of the population, the exact definitions of health states within the economic model, the timing of data collection, sample size, and mode of administration. Missing data can be an issue, particularly in longitudinal studies and it is important to determine if the missing data will bias inferences from analyses. For example respondents may fail to complete follow-up questionnaires because of a relapse or severity of their condition.

The decision on the preferred study type and the particular quality of life measure should be informed by any evidence currently available in the literature, the design of data collection, and the exact requirements of the model that will be used to support resource allocation decisions (e.g. reimbursement).

KEY POINTS FOR DECISION MAKERS

- A literature review and a conceptual model are recommended early within a R&D programme to help identify the required health state utility values.
- While clinical trials can provide an opportunity to collect the required evidence, their appropriateness should be reviewed against the requirements of the model (population characteristics, time horizon, and sample size) and alternative sources (observational studies, registries) may be more appropriate.
- Considerations when designing the collection of the evidence include: patient and investigator burden, whom to ask, the representativeness of the population, the exact definitions of health states within the economic model, the timing of data collection, sample size, and mode of administration.

1. Introduction

When considering health state utility evidence for use in decision models, the recommended starting point is early within a R&D programme, and before the development of any trial protocol [1]. An initial literature review is essential to determine if the required health state utility values (HSUVs) are currently available or if new evidence is required [2,3]. A conceptual model framework can help to identify the requirements of the decision analytic model (DAM) and a modeller should be consulted to determine the requirements prior to conducting the literature review, or designing protocols for any associated clinical trials or observational studies [1]. However, DAMs regularly evolve over time and it is possible that the full requirements will not be known until the final health states have been agreed. Consequently while an early or conceptual model can inform trial protocols, the final DAM is more useful when examining the literature.

This article covers the advantages and disadvantages of collecting preference-based measures and other health related quality of life measures to generate HSUVs in alternative study designs (e.g. randomised control trials (RCTs), observational studies, registries), considerations when designing the collection of evidence (e.g. relevance of the population, timing of collection, mode of administration), and recommendations for data analysis and reporting standards. The focus is the use of clinical studies to collect HSUVs for use in DAMs designed to compare the long-term benefits of alternative interventions (as opposed to economic evaluations conducted alongside short-term clinical trials).

2. Alternative sources of evidence for HSUVs

There are numerous forms of study design that can be used to collect HSUV evidence including clinical trials (defined as randomised clinical studies comparing the efficacy of interventions), observational studies such as cross-sectional (collect evidence at one point in time) or longitudinal (collect repeated observations of the same variable over periods of time), registries (generally longitudinal, and collect specified outcomes from a particular population generally defined by a specific condition such as cancer or cardiovascular disease). There is also the option of vignette studies, bespoke studies designed to elicit preferences for a small number of predefined health states (see Brazier et al. for more discussion on this approach [4]).

2.1 Clinical trials

Clinical trials can provide important and useful means to collect the required HSUVs. Their use preserves and enhances internal validity as the benefits (clinical effect and HSUVs) of the interventions are collected within the same study and the HSUVs are collected directly from recipients of the intervention(s) under appraisal. This has the additional benefit of being able to measure the impact of side-effects of the interventions. There are however, numerous limitations such as the representativeness of the study population, the enhanced medical care provided, the sample size, time horizon of the trial, and potential implications for regulatory approval.

Representativeness of the study population

Strict exclusion criteria such as exclusion of younger or older ages, disease severity or people with comorbidities, on non-study medications, with a recent history of related clinical events (e.g. immediately post stroke) or at the extremes of the disease spectrum can mean that the trial population may not be fully representative of the target population in clinical practice. While adjustments may be made using prognostic models to account for some of these factors, this would introduce an additional level of uncertainty and is less credible than collecting evidence on the full range of individuals of interest.

Enhanced medical care

Clinical protocols may involve high levels of monitoring and follow-up, and investigational procedures, not generally observed in routine clinical practice. It is possible that the levels of healthcare provided may increase recipients' sense of well-being and potentially health related quality of life (HRQoL), thus decreasing confidence in generalisability to that observed in patients in routine clinical practice.

Trial time horizon

Clinical trials tend to use relatively short time horizons whereas DAMs frequently need the lifetime health benefits and costs associated with interventions. Consequently, HSUVs collected in clinical trials may not be suitable for all health states within the DAM and are generally not useful for longer-term evidence (e.g. how HSUVs change over time after the fractures) or specific clinical events as trials are designed to compare arms rather than the effect associated with clinical events. The time horizon also has implications when HSUVs are required for rarer, less frequent events as these may not be observed within the short time horizon of the clinical trial. Alternative or additional sources such as registries or observational studies may provide more appropriate evidence in these instances.

Sample size

Study sample size for experimental studies is generally calculated to detect a difference (or equivalence) in clinical effect rather than HRQoL. Consequently, when subgrouping by clinical event type (e.g. fracture site in osteoarthritis trials), the number of patients within the trial who experience a specific event may be too small to detect a statistically significant difference in HSUVs. While this may not be strictly necessary for an economic model, smaller samples will increase the uncertainty in results. In addition, if the data are analysed by event type and treatment arm, any observed difference in mean HSUV may not be detected as statistically significant as the difference is too subtle to detect, these differences may be further effected by factors such as missing data, cross-over and loss to follow-up (see example in **Box 1**).

Implications for regulatory approval

The use of generic preference-based measures to collect HSUVs in trials can be misleading where it has been powered on different endpoints {O'Brian classic paper – jeb to find the reference]. This carries a risk as if the FDA/EMA review thes evidence and the results are not statistically significant, or are inconsistent with other patient reported outcome measures, this may affect the chances of approval.

The value of patient reported outcome data to the intervention under appraisal should be assessed to determine if this is likely to be a key differentiator. The GPBM should be examined to determine if it could capture the treatment effect in a given population. One way around this would be to include the PBM as an exploratory endpoint and only include descriptive data (e.g. responses to the health dimensions) in the statistical analysis plan, as opposed to using the traditional approach of comparison between treatment arms. Depending on the HSUVs required to satisfy the DAM, one alternative could be to limit the collection of the GPBM to the baseline and clinical events (e.g. immediately post hip fractures) only, as opposed to end of follow-up. One could also argue that comparison across the trial arms could be misleading due to censoring and loss to follow-up etc.

Box 1 Example from the UKPDS (Clarke et al, 2002 [5])

The landmark UK Prospective Diabetes Study (UKPDS) of glycaemic therapies in patients with newly diagnosed type 2 diabetes demonstrated that intensive blood glucose control significantly reduced the long-term sequelae (micro and macrovascular complications) of diabetes.

While, no difference in EQ-5D evidence (n = 3,667) was observed across treatment arm when using conventional significance levels, subsequent analyses subgrouped long-term clinical outcomes (myocardial infarction, ischaemic heart disease, stroke, heart failure, amputation and visual acuity) clearly demonstrated a significant detrimental effect on HSUVs associated with the events [5,6] .

2.2 Observational studies

There are numerous reasons why observational studies may be preferred over clinical trials including the arguments that evidence from clinical trials is not representative of general clinical practice due to the exacting inclusion and exclusion criteria, the more intensive care provided to the participants, the time horizon and the regimented timing of data collection. Bespoke observational studies can provide extremely important 'real-world' utility evidence and can range from simple online surveys to complex longitudinal surveys with repeated measures collected over time. Observational study designs can be cross-sectional (collect evidence at one point in time) or longitudinal (collect repeated observations of the same variable over periods of time).

Longitudinal studies can be designed to examine the immediate and long term effects associated with discrete events (such as the acute period after a fracture or the longer-term after rehabilitation), or they can collect changes over time associated with chronic progressive conditions such as rheumatoid arthritis or Crohn's disease. Prospective cohort studies (a sub-type of longitudinal studies) can be designed to target a particular subgroup of patients matching the characteristics required for rare health states or clinical events. The main disadvantages of prospective studies are the resource (both cost and time) implications involved in identifying and recruiting the required sample.

Cross-sectional surveys provide evidence collected at one point in time and thus there may be difficulties in matching evidence of HRQoL with precisely defined health states. For example a myocardial infarction could be defined as the immediate period following the acute clinical event, or later periods such as 6 months, 12 months or five years after the event. While it is possible to subgroup by time since event in these cases, a much larger sample size is required. However, compared to longitudinal studies, they are a relatively quick and inexpensive source of evidence.

In addition to the costs of conducting a bespoke study , there can be problems with recruitment (e.g. when the participants' characteristics are to match some pre-defined clinical definition such as severity or history of events), and retention (particularly when conducting longitudinal studies). Patients' medical history (i.e. current or previous health conditions) may be self-reported which is not always considered to be reliable [7].

2.3 Registries

Registries can provide an alternative source of evidence [8,9,10,11,12]. The benefits are the comparatively low costs associated with obtaining the data (if owned by a second party), and they may also include patients eligible for the intervention in clinical practice who are excluded from the clinical trials. However, when using existing datasets designed to satisfy other research questions, there may be problems in matching the requirements of the DAM with the data available.

2.4 Reviews of the literature

A literature review is a requirement of some Health Technology Assessment (HTA) agencies (e.g. the National Institute for Health and Care Excellence NICE) [13] and the results of this can be informative in terms of both the measure to be used to collect the HSUVs and the range of estimates that should be explored in economic sensitivity analyses [14].

2.5 Supplementary evidence

Finally, consideration should be given to including preference-based measures in phase 4 studies. These studies may include real-world evidence used in clinical practice and tend to be longer in duration than phase 3. They may use less stringent inclusion/exclusion criteria thus providing greater opportunities to obtain the evidence required for re-assessment and future reimbursement submissions, and can be used to compare with and support the evidence collected in the earlier clinical trials.

3. What to consider when designing the collection of data

There are a range of issues to consider when designing the collection of HSUVs irrespective of the study design or measure selected. These include matching the definitions of the anticipated health states/events in the early economic model, whom to ask, the representativeness of the study population, the timing of assessments, repeated measures, mode of administration, sample size, and patient and investigator burden.

3.1 Whom to ask

It is common practice to obtain health status descriptions directly from patients as they are generally in the best position to know how their health has affected their function and well-being [15,16]. Responses from proxies (e.g. family member, principle carer, clinician etc.) are not always directly comparable with those obtained from patients, regardless of the measure used (see example in **Box 2**) [17,18]. Consequently if used in a DAM, this potential bias should be acknowledged in the text and a series of sensitivity analyses conducted to illustrate the effect of the uncertainty in this evidence.

However, there are instances when it is not possible to ask patients to rate their own health such as when patients have severe mental health conditions or are too ill, and young children. In these cases, responses from proxies may be used but it is recommended that patient responses are used when they are willing and able to provide this evidence [13,16].

Box 2 Patient versus proxy measurement of health related quality of life in dementia

EQ-5D: Patients' carers reported higher levels of disability than the patients across all five dimensions on the EQ-5D in a study of patients with dementia [18]. Conversely, clinicians reported fewer problems on the dimensions 'pain/discomfort' and 'anxiety/depression'. The level of agreement between responses was assessed as only fair.

ICECAP: Work experience and gender were reported to influence proxy responses to the ICEpop CAPability Measure for Older People (ICECAP-O) when used to assess the well-being in older patients with dementia residing in a nursing home [19].

There is an additional consideration where there is a significant impact on the informal carer(s), such as the case with dementia or Parkinson's [20]. To measure the full impact of an intervention it will be necessary to collect HRQoL data from the carer, in addition to the person taking the intervention that they take care of.

3.2 Representativeness of the study population

The study participants should reflect the population in the DAM i.e. patients who would receive the interventions under appraisal if it was provided in routine clinical practice. In many cases, if a particular subgroup in the DAM has been excluded from the clinical trial, evidence from observational studies may be preferable as these studies may be designed to recruit patients with pre-defined characteristics from the target population.

3.3 Fit the decision problem: the definitions of health states within the decision analytic model

Care should be taken to ensure the characteristics of the target population cover the full range of definitions (health states, events) of health states within the early DAM. This is particularly relevant for progressive conditions (e.g. arthritis or Parkinson disease where patients at the more severe end of the disease spectrum may be less likely to respond), or when evidence from relatively rare events or complications is needed. In the latter instance, a prospective study targeting a specific subgroup more likely to experience the event may be considered. However, responses rates are likely to be extremely poor in patients experiencing serious adverse events or approaching end of life in palliative care [1]. In these cases, evidence from proxies should be considered (see **sub-Section 3.1**).

3.4 Timing of data collection

Assessments in clinical trials are generally at scheduled intervals such as when administering the intervention at out-patient appointments (e.g. 1 week, 4 week, 12 weeks). These time-points may not coincide with the desired timing to capture the effects of interventions and clinical endpoints required for the DAM such as symptom flares or hospitalisations. For example, HSUVs in cancer trials are generally collected during visits for chemotherapy (approximately every three weeks) but these are unlikely to capture the effects of chemotherapy toxicity that generally occur between treatments, and start just after disease progression or the end-of-life period. Similarly, inflammatory conditions are characterised by flares in symptoms, and it could be difficult to capture the HSUVs associated with these if the scheduled intervals are adhered to. The schedule for collection of utility data should be flexible to enable the capture of changes associated with such flares (or discrete clinical events) and should be synchronised with the collection of condition specific measures.

For DAMS where the clinical variable used to describe the effectiveness of an intervention is used to represent progression over time, such as arthritis (uses the Health Assessment Questionnaire (HAQ)) or ankylosing spondylitis (uses the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) and the Bath Ankylosing Spondylitis Functional Index (BASFI)) it is important to ensure that the clinical variable is collected at the same time-point as the evidence for the HSUV. Models used to evaluate interventions in these types of chronic progressive conditions tend to use a mapping function that describes the relationship between the clinical variable and the HSUV to interpolate and extrapolate the required HSUVs across the full disease spectrum [21].

DAMs frequently use a life time horizon and depending on the condition of interest, they may require HSUVs that change over time such as in progressive conditions. As evidence is rarely collected in clinical trials for a sufficient length of time,, historically, analysts have sourced this form of evidence from cross-sectional studies as these are less expensive and easier to collect (due to problems with non-response/retention) than repeated-measure studies. However, there is evidence suggesting that cross-sectional data may provide biased estimates of the effects on HRQoL as differences in observed quality of life between patients with and without an event may be due to underlying heterogeneity across the two types of patients rather than due to the event [6]. This bias is introduced as cross-sectional analyses attribute all observed differences in quality of life between patients with and without an event to the event, when in reality some or all the difference may be due to underlying heterogeneity across the two types of patients rather than the event.

Conversely, missing data is common in longitudinal data and can be extremely problematic due to the non-random nature of drop-outs. This can contribute to bias if the missing data is systematically different from the observed evidence [22].

In addition to matching the requirements of the DAM, when scheduling the timing of data collection, the recall period of the measure should be considered. The EQ-5D for example asks respondents to value their health today whereas the SF-36 and SF12 use recall periods of 4 weeks and 1 week (acute version) [23,24]. For example, the recall period of the EQ-5D may be more appropriate than the SF-36 in an efficacy study of clopidogrel in patients undergoing a cardiac stenting procedure when the HSUVs for the DAM are required for the days immediately after surgery. Conversely, if the evidence collection points are pre-scheduled to match clinical appointments in an efficacy study, the recall period of the SF-36 may be more appropriate than that of the EQ-5D in conditions characterised by flares, as the probability of having a flare on the pre-scheduled day is smaller than the probability of having a flare between visits.

3.5 Sample size

As the objective is to collect HSUVs for use in DAMs, the sample size should be governed by the need for precision (and uncertainty) in the HSUVs, rather than statistical comparisons between arms within the study. If subgroup analyses are to be conducted (e.g. examining differences in HSUVs for numerous discrete clinical events such as in a cardiovascular disease study), this should be considered in the sample size calculation. In DAMs sensitive to the HSUVs, the sample size required to reduce uncertainty around the point estimate may be calculated using value of information techniques, otherwise standard techniques using confidence intervals [16]. Value of information techniques consider the uncertainty surrounding the current evidence and the implications of reducing the uncertainty around the incremental cost-effectiveness ratio for a given threshold in a DAM and thus the added value to a decision maker (in this case through increasing the sample size for the HSUVs) [16].

3.6 Mode of administration

The vast majority of HRQoL measures like GPBMs are designed as self-administered questionnaires to be completed using pen/paper, online or via tablets and smartphones. The main advantages of self-completion are the relatively low cost and increasing the probability of acquiring responses to sensitive questions. The main disadvantages are the low response rates and the difficulties in obtaining a truly accurate sample representative of the full target population which may introduce responder bias (where respondents differ from non-respondents in terms of socio-demographic and other characteristics). Electronic versions are now available for all the main GPBMs and a recent study comparing evidence on the EQ-5D-5L collected using either paper or a mobile phone app reported equivalent results and response rates [25,26].

Evidence can also be collected using interviews where the interviewer rather than the respondent records the responses. Face-to-face interviews tend to be used where additional and often complex information is required above the evidence on HRQoL. Although face-to-face interviews are an expensive method of collecting evidence and introduce another potential source of bias from the (interviewer), response rates are typically higher than for postal surveys and as information is obtained from target respondents, sample composition bias is generally reduced. Telephone interviews are a relatively quick and low cost alternative compared to face-to-face but response rates tend to be lower than for face-to-face interviews. The advantage over face-to-face interviews is that interviewer bias can be reduced through close supervision, and a more accurate assessment is obtained when patients are at home rather than in the clinic [27]. The disadvantage is that respondents may find complex questions difficult to answer, there is still the risk of interviewer bias, and while lower than face to face interviews the cost is still higher than web-based collection. There is some evidence suggesting that the mode of administration can affect responses, for example respondents may be more likely to report the highest or the lowest categories when responding to an oral rather than a written self-report health questionnaire [28].

Due to differences from mode and who administers the questionnaire, such as sample representativeness, response rates, comprehension and response strategies, it has been suggested that the mode of administration is standardised (i.e. that only one mode of administration is used within a study) [1]. However, choice depends on factors such as the characteristics of the target population (e.g. response rates could be low when using electronic versions in elderly populations), and time or resource constraints and the ultimate decision depends on the relevance of the mode of administration to the individual study. If more than one mode of administration is required to optimise the flexibility and thus response rates (e.g. in rare conditions), the collection mode could be used as a covariate in subsequent analyses and the modes of collection agreed apriori.

3.7 Patient and investigator burden

The inclusion of a GPBM or other HRQoL measure in a clinical trial is commonly objected to on the grounds of the additional burden placed on the patient and investigator [29]. This can be particularly relevant when frequent assessments are required to satisfy the needs of the DAM. The requirements of the differing authorities should be balanced and the importance of high quality evidence on HSUVs to reimbursement authorities should be considered, particularly when it is necessary to synchronise data collection timing with clinical variables or tests.

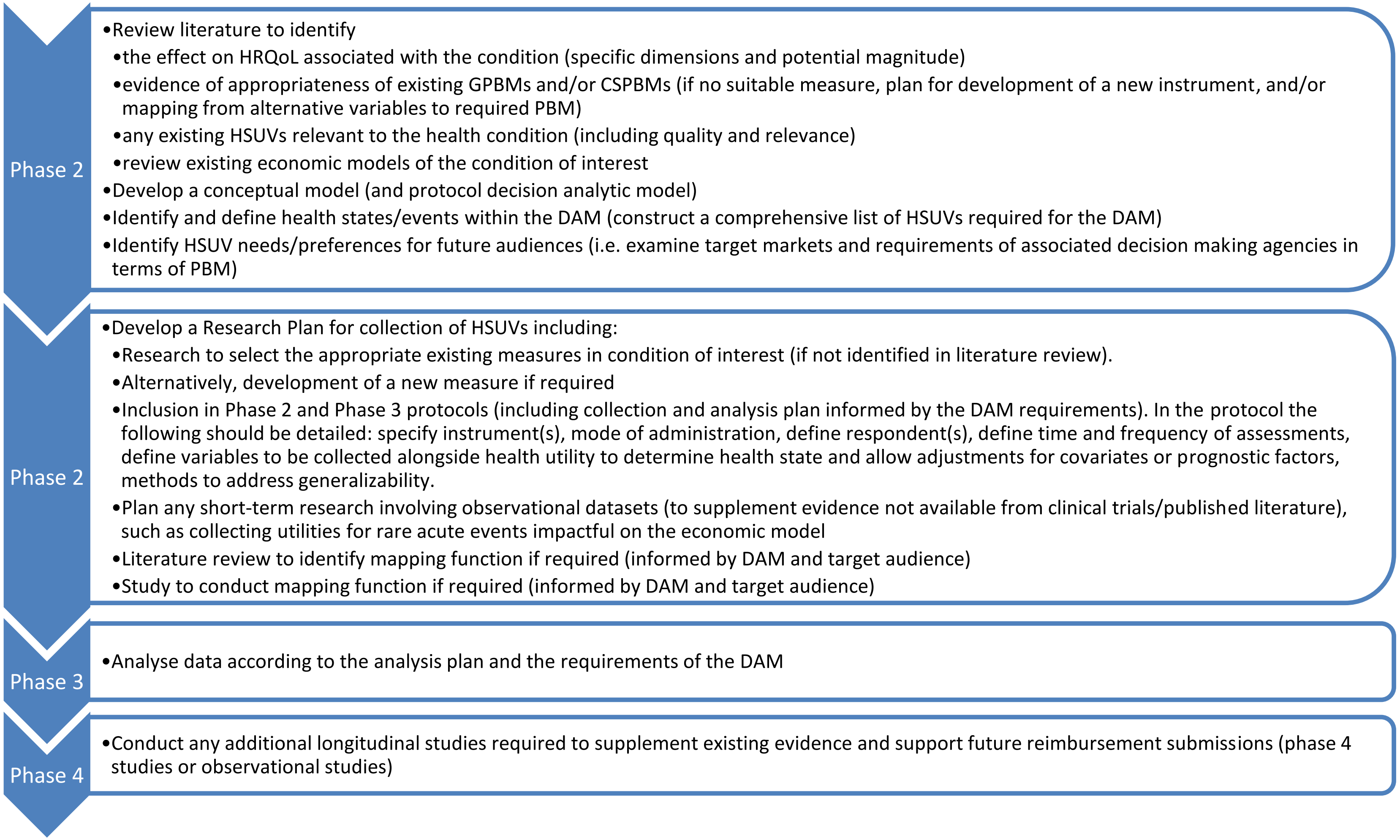
Some measures (e.g. the EQ-5D) take minutes to complete, while others can be more burdensome [30,31]. Ultimately HSUVs should be collected using a measure that is appropriate in the health condition of interest, and satisfy the relevant reimbursement authority [32]. If the psychometric properties of the required measure have not been assessed in the target population, inclusion in phase II studies could provide an opportunity to address this.

4. Recommendations for developing HSUV research plan

To ensure that appropriate HSUV evidence is available when needed it is important to develop and document an early HSUV research plan/strategy (see **Figure 1**). An early understanding of the exact requirements of the DAM, and continued input from an experienced healthcare modeller is essential at all stages to ensure the required evidence is available when it is needed. The strategy for collection of HSUVs should be developed in conjunction with the intervention's R&D programme, and should be informed by a literature review and at the very least an early conceptual economic model during phase 1 and 2 clinical studies. The conceptual model, an often overlooked stage in research and development plans, will assist in the development process of the actual decision analytical model and will help to identify the HSUV evidence needed. Ideally, a protocol DAM should also be constructed very early (during phase 2 clinical studies).

Phase 4 studies may provide the opportunity to collect additional longitudinal real-world evidence to supplement and compare with the existing evidence and can be used to support reassessment and re-submissions at later dates.

Figure 1 Tasks at each phase of clinical studies (from phase 2)



Phase 2

- Review literature to identify
  - the effect on HRQoL associated with the condition (specific dimensions and potential magnitude)
  - evidence of appropriateness of existing GPBMs and/or CSPBMs (if no suitable measure, plan for development of a new instrument, and/or mapping from alternative variables to required PBM)
  - any existing HSUVs relevant to the health condition (including quality and relevance)
  - review existing economic models of the condition of interest
- Develop a conceptual model (and protocol decision analytic model)
- Identify and define health states/events within the DAM (construct a comprehensive list of HSUVs required for the DAM)
- Identify HSUV needs/preferences for future audiences (i.e. examine target markets and requirements of associated decision making agencies in terms of PBM)

Phase 2

- Develop a Research Plan for collection of HSUVs including:
  - Research to select the appropriate existing measures in condition of interest (if not identified in literature review).
  - Alternatively, development of a new measure if required
  - Inclusion in Phase 2 and Phase 3 protocols (including collection and analysis plan informed by the DAM requirements). In the protocol the following should be detailed: specify instrument(s), mode of administration, define respondent(s), define time and frequency of assessments, define variables to be collected alongside health utility to determine health state and allow adjustments for covariates or prognostic factors, methods to address generalizability.
  - Plan any short-term research involving observational datasets (to supplement evidence not available from clinical trials/published literature), such as collecting utilities for rare acute events impactful on the economic model
  - Literature review to identify mapping function if required (informed by DAM and target audience)
  - Study to conduct mapping function if required (informed by DAM and target audience)

Phase 3

- Analyse data according to the analysis plan and the requirements of the DAM

Phase 4

- Conduct any additional longitudinal studies required to supplement existing evidence and support future reimbursement submissions (phase 4 studies or observational studies)

Source: Adapted from Wolowacz et al 2016 [1]. **Key:** CSPBM – condition-specific preference-based measure; DAM – decision allocation model; GPBM – generic preference-based measure; HRQoL – health related quality of life; HSUV – health state utility value; PBM – preference-based measure

5. Recommendations for data analysis

The statistical analysis plan should not be constrained by the traditional approach used when analysing data for regulatory purposes (i.e. comparison between treatment arms) but should satisfy the needs of the economic analysis [1]. Country specific preference-weights/tariff relevant to the DAM should be applied to data from all countries for multinational studies. A recent systematic review of economic evaluations alongside multinational studies showed that methods of analysis were inconsistent between studies [33]. However, country covariates, derived from statistical regression models such as multi-level models, in order to allow for clustering, may be used to adjust for any differences in HRQoL for multinational studies. The value of the evidence to future economic evaluations may be maximised through the inclusion of prognostic factors as covariates in statistical modelling to enable the results from a clinical study to be adjusted to reflect the characteristics of populations in routine clinical practice. If statistical regression models are generated and the results used to predict the HSUVs in a DAM, the associated covariance should be reported to enable the integrity of the ordering of the HSUVs to be retained under conditions of uncertainty [34].

Missing data can be problematic for HRQoL evidence as patients may be assessed at several time points. If data are missing it is important to establish the level of missing data and the type of missingness. Data may be missing completely at random, missing at random, or missing not at random also known as non-ignorable missingness [35]. Understanding why data are missing is important for HRQoL evidence, particularly if the population are severely ill as a proportion of respondents may fail to complete the measure because of the severity of their condition (or death). These missing data are non-ignorable as they are directly dependent on the health status of the patient, and in these cases it is inappropriate to analyse the data using complete cases only (i.e. dropping respondents with missing data from the analyses

Methods used to handle missing data range from simple approaches whereby patients whose data is incomplete is deleted, to more complex approaches, such as multiple imputation [36], where missing data points are imputed in some way. The former is advantageous in terms of simplicity and ease of analyses, but in some cases the sample size can reduce substantially losing statistical power, further the variance in the estimates can be underestimated [37]. Not all the information is used, and unless the data is missing completely at random, the results are likely to be biased. The main imputation methods include single imputation (mean/mode substitution, dummy variable method, single regression), which can still underestimate the variance [See e.g. Briggs, Rubin, Schafler] and modelbased methods (maximum likelihood, multiple imputation) [36,37,39]. These methods work well if data are either missing at random or missing completely at random, if data are missing not at random them methods such as selection models and pattern mixture models should be used as these allow for systematic missingness (See for example Laird [38]). A review of methods used to handle missing data in economic evaluations conducted alongside clinical trials, reported that complete case analysis was the most common methodology, and frequency had increased over time despite the introduction of more sophisticated methods [39].

6. Recommendations for reporting

The foremost recommendation on reporting standards for evidence on HSUVs is transparency and justification from the initial choice of source of evidence through to the presentation of results. Any reviews conducted to inform the collection of additional evidence should be duly reported (04 Ara 2017). In addition to the usual information on study design, sample size, and summary variables describing the study population (age, gender, health status etc), it is important to provide information on the full range of the variables used within the analyses. This will allow reviewers to assess the relevance of the data in terms of where it will be used. The evidence should be compared directly with the characteristics of the patients within the DAM. Any subgroup analysis such as age or disease severity should be clearly explained together with the size.

7. Summary

While clinical trials can provide an efficient method of collecting the required HSUVs, observational studies provide a useful alternative when it is not possible (or desirable) to collect the HSUVs in clinical trials. The decision on the preferred study type and the PBM measure should be informed by any evidence currently available in the literature, the design of data collection, and the exact requirements of the DAM that will be used to support reimbursement.

Disclosure statement

This article is published in a special edition journal supplement wholly funded by Takeda Pharmaceutical International AG, Zurich, Switzerland.

Acknowledgement

The authors would like to thank Prof Jon Karnon, PhD of The University of Adelaide and Dr Andrew Lloyd, Phd of Bladen Associates Ltd for their editorial review.

Author contributions

RA reviewed the literature, wrote the first, subsequent and final drafts of the manuscript. JEB made significant edits to the first and final draft of the manuscript.

Compliance with Ethical Standards

Funding This study was funded by an unrestricted grant from Takeda Pharmeceuticals International AG.

Conflict of interest Roberta Ara has no conflicts of interest. John Brazier has no conflicts of interest. Tracey Young has no conflicts of interest.

REFERENCES

[1] Wolowacz SE, Briggs A, Belozeroff V, et al. Estimating health-state utility for economic models in clinical studies: an ISPOR good research practices task force report. Value in Health. 2016; 19(6):704-719.

[2] Papaioannou D, Brazier J, Paisley S. Systematic searching and selection of health state utility values from the literature. Value in Health. 2013;16:686-95.

[3] Papaioannou D, Brazier J, Paisley S. The identification, review and synthesis of health state utility values from the literature. NICE DSU Technical Support Document. 2010;9.

4 Brazier JE, Rowen D. NICE DSU Technical Support Document 11: Alternatives to EQ-5D for generating health state utility values. 2011. Available fromhttp://www.nicedsu.org.uk accessed 17th March 2017.

[5] Clarke P, Gray A, Holman R. Estimating utility values for health states of type 2 diabetic patients using the EQ-5D. Medical Decision Making 2002, 22: 340–349.

[6] Alva M, Gray A, Mihaylova B, Clarke P. The Effect of Diabetes Complications on Health-Related Quality of Life: The Importance of Longitudinal Data to Address Patient Heterogeneity. Health Economics 2014, 23(4): 487-500.

[7] Smith B, Chu LK, Smith TC, Amoroso PJ, Boyko EJ, Hooper TI, Gackstetter GD, Ryan MA. Challenges of self-reported medical conditions and electronic medical records among members of a large military cohort. BMC Medical Research Methodology. 2008 Jun 5;8(1):37.

[8] Zink A Listing J Kary S Ramlau Pet al. Treatment continuation in patients receiving biological agents or conventional DMARD therapy. Ann Rheum Dis 2005;64:1274–9.

[9] Lindqvist UR, Alenius GM, Husmark T, et al. The Swedish early psoriatic arthritis register—2-year followup: a comparison with early rheumatoid arthritis. J Rheumatol 2008;35:668–73.

[10] Kirchhof P, Ammentorp B, Darius H *et al*. Management of atrial fibrillation in seven European countries after the publication of the 2010 ESC Guidelines on atrial fibrillation: primary results of the PREvention oF thromboembolic events–European Registry in Atrial Fibrillation (PREFER in AF). *Europace* 2014;16:6–14.

[11] Agnelli G, Gitt AK, Bauersachs R, Fronk EM, Laeis P, Mismetti P, Monreal M, Willich SN, Wolf WP, Cohen AT. The management of acute venous thromboembolism in clinical practice–study rationale and protocol of the European PREFER in VTE Registry. Thrombosis journal. 2015 Oct 21;13(1):41.

[12] Jones KH, Ford DV, Jones PA, John A, Middleton RM, Lockhart-Jones H, Peng J, Osborne LA, Noble JG How People with Multiple Sclerosis Rate Their Quality of Life: An EQ-5D Survey via the UK MS Register. PLoS One. 2013 Jun 11;8(6):e65640. doi: 10.1371/journal.pone.0065640. Print 2013.

[13] NICE (National Institute for Health and Care Excellence). Guide to the methods of technology appraisal. London: National Health Service, 2013. Available from www.nice.org.uk

[14] Ara R, Brazier J, Peasgood T, Paisley S. The identification, review and synthesis of HSUVs from the literature. Current issue Pharmacoeconomics.

[15] Bowling A. Mode of questionnaire administration can have serious effects on data quality. Journal of Public Health 2005, 27(3): 281-291.

[16] Brazier J, Ratcliffe J, Tsuchiya A, Solomon J, editors. Measuring and Valuing Health for Economic Evaluation. Oxford: Oxford University Press; 2nd edition 2017.

[17] Eiser C, Varni JW. Health-related quality of life and symptom reporting: similarities and differences between children and their parents. European Journal of Pediatrics 2013, 172: 1299-1304.

[18] Coucill W, Bryan S, Bentham P, Buckley A, Laight A. EQ-5D in patients with dementia: an investigation of inter-rater agreement. Medical Care 2001, 39: 760–1.

[19] Makai P, Beckebans F, van Exel J, Brouwer WB. Quality of Life of Nursing Home Residents with Dementia: Validation of the German Version of the ICECAP-O. PLoS ONE 2014, 9(3): e92016.

[20] Al-Janabi H, Van Exel J, Brouwer W, Coast J. A framework for including family health spillovers in economic evaluation. Medical Decision Making. 2016 Feb;36(2):176-86.

[21] 05 Ara R, Rowen D, Mukuria C. The use of mapping to estimate health state utility values, Current issue Pharmacoeconomics.

[22] Fayers PM, Machin D. Modelling Longitudinal Data. Quality of Life: Assessment, Analysis and Interpretation.:203-23.

[23] Ware Jr JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. Medical care. 1992 Jun 1:473-83.

[24] Ware Jr JE, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Medical care. 1996 Mar 1;34(3):220-33.

[25] Mulhern B, Longworth L, Brazier J, Rowen D, Bansback N, Devlin N, Tsuchiya A. Binary choice health state valuation and mode of administration: head-to-head comparison of online and CAPI. Value in Health. 2013 Feb 28;16(1):104-13.

[26] O'Gorman H, Mulhern B, Rotherham N, Brazier J. Comparing the equivalence of EQ-5D-5L across different modes of administration. Value in Health 2014, 17: A517.

[27] Lyons RA, Wareham K, Lucas M, et al. SF-36 scores vary by method of administration: implications for study design. J Pub Health Med 1999; 21: 41–45

[28] Clarke PM, Ryan C. Self-reported health: reliability and consequences for health inequality measurement. Health Economics 2006, 15(6): 645-652.

[29] Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: is shorter better? a review and meta-analysis. Value in Health. 2011 Dec 31;14(8):1101-8.

[30] 02 Brazier J, Ara R, Rowen D, Chevrou-Severac H. A review of Generic preference-based measures. Current issue Pharmacoeconomics.

[31] 03 Rowen D, Brazier J, Ara R, Azzabi Zouraq I, The role of condition-specific preference-based measures. Current issue Pharmacoeconomics.

[32] 01 Rowen D, Azzabi Zouraq I, Chevrou-Severac H, van Hout B. International regulations and recommendations Current issue Pharmacoeconomics.

[33] Oppong R, Jowett S, Roberts TE. Economic evaluation alongside multinational studies: a systematic review of empirical studies. PloS one. 2015 Jun 29;10(6):e0131949.

[34] Wailoo AJ, Hernandez-Alava M, Manca A, Mejia A, Ray J, Crawford B, Botteman M, Busschbach J. Mapping to Estimate Health-State Utility from Non–Preference-Based Outcome Measures: An ISPOR Good Practices for Outcomes Research Task Force Report. Value in Health. 2017 Jan 31;20(1):18-27.

[35] Polit DF Beck CT (2012). *Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th ed*. Philadelphia, USA: Wolters Klower Health, Lippincott Williams & Wilkins.

[36] Little & Rubin Statistical Analysis with Missing Data. John Wiley & Sons, New York, 1987

[37] Briggs A, Clark T, Wolstenholme J, Clarke P (2003) Health Economics: 12; 377-392. Missing…. Presumed at random. http://onlinelibrary.wiley.com/doi/10.1002/hec.766/epdf

[38] Laird NM. Missing data in longitudinal studies. Statistics in medicine. 1988 Jan 1;7(1-2):305-15.

[39] Schafer JL. Analysis of incomplete multivariate data. CRC press; 1997 Aug 1.