Work in progress – do not quote.

Collège des Économistes de la Santé / Health Economists Study Group
Joint Workshop   Paris, January 14-16 2004

Quality weighting in evidence synthesis for decision support is a matter of preferences – QE-5D is proposed as an instrument for their elicitation.

Jack Dowie and Zaid Chalabi
London School of Hygiene and Tropical Medicine

While everyone appears to be broadly in favour of assessing the 'quality' of evidence in a serious way, controversy has surrounded the 'quality weighting' of evidential sources, not only in meta-analyses of studies employing the same basic study design (e.g. randomised controlled trials; cohort, case control and other non-randomised designs) but especially in wider cross-design research syntheses, where the sources go beyond those just mentioned to include such things as registry data and expert opinions.

This paper has only a derivative interest in the debate on quality weighting in relation to *knowledge* technologies, where the aim is to establish what we *know*, by criteria and standards set by statistical science for science. Insofar as 'knowledge technologists' seek to contribute to decision making they typically see their function as being to hand over to decision makers information which has been scientifically and statistically screened - as what we will call *Knowledge* Support (i.e. support in the form of 'synthesised' knowledge). The decision makers are assumed, not only to have the power and right to make decisions (which by definition they do have), but also to have the capacities to deal satisfactorily with all the demands of decision making, including the problems of evidence synthesis that have been left unresolved by the knowledge technologists - necessarily so because they *are* currently unresolvable by scientific standards. Among these tasks will be the quality weighting of either whole studies, or of the quality assessments for individual components of each study, if a multi-item checklist/profile is what the knowledge technologist has seen it proper to supply.

At the moment decision and policy makers almost universally employ a predominantly *intuitive* decision technology, one that we will refer to as 'taking into account and bearing in mind' (TIABIM for short). Among the things they must somehow 'take into account and bear in mind' are the quality variations in evidential sources as reported in the knowledge support they receive. Our interest in quality weighting arises in the context of commitment to a very much more *analytical* decision technology, where Bayesian Decision Analysis is undertaken and offered to decision makers not as Knowledge Support but as *Decision* Support - as a means of *making* the decision. An alternative means to the TIABIM technology which they now employ. The decision makers will, as already noted, have both the right and the power to reject the option identified by any particular instantiation of this alternative Decision Technology (DT), but it is important to see that, in the absence of formal comparative evaluation, their acceptance or rejection has no implications regarding the relative quality - or indeed ethicality - of the instantiations of the two DTs. (Dowie, 2001)

The comparative evaluation of instantiations of alternative DTs is not, however, a concern of this paper. It simply suggests (a) that there is no alternative to quality weighting *across all evidential sources if* one is employing a Decision Analytic DT, as opposed to a TIABIM one, and (b) that quality weighting, in whichever context it is undertaken, needs to be interpreted as a matter of (subjective) *preferences* as well as of (subjective) beliefs. The former proposition is likely to be controversial only within Decision Analytic circles. The latter will, on our reading of the literature, be much more profoundly and widely controversial.

In the first section we selectively survey the literature on quality weighting in meta-analyses for Knowledge Support, in order to establish the main sorts of argument that surround it. In section two we look at approaches that are more inclined towards the idea of providing Decision Support to decision makers - the Confidence Profile Method (CPM) and its various successors in the form of 'Bayesian comprehensive decision modelling'. In section three we state the case for explicit quality weighting of evidence from across the entire spectrum of evidential sources when deriving estimates of parameters or parameter distributions for Bayesian Decision Analyses. It is proposed that, assuming all statistical de-biasing and adjusting that is feasible and possible within the decisional time window have been undertaken, quality weighting of the various parameters should be based on a *descriptive* encoding of the types of bias exhibited by each study and their magnitudes (this encoding reflecting the subjective *beliefs* of experts), to which is applied an *evaluative* generic KRQOE (Knowledge-Related Quality of Evidence) tariff that reflects relevant subjective *preferences* in relation to the types and magnitudes of bias covered by the classification system.

Given the need to actually elicit these preferences (as opposed to suggest some ungrounded and essentially arbitrary ones), the classification matrix must be practical and decision-requisite in terms of dimensions and levels. It cannot be of the complexity demanded by Knowledge Technology criteria and standards. This situation is closely analogous to that in HRQOL measurement and we propose an instrument, QE-5D, on the basis of which (a) the required generic encoding and (b) the preference-based weighting of evidence across all study designs and evidential sources, can be undertaken.

Our main hope in advancing this instrument as a tentative proposal for discussion among interested parties at an early stage, is to try to avoid the messy situation that has arisen in HRQOL, where we have a number of competing generic indexes, including EQ-5D (and -6D), HUI, QWB, and AUSQOL. A second hope is to prevent the development of (a) any more Quality *scales* that lack explicit grounding in preferences or are too complex to be so grounded, and (b) any more Quality *checklists* that lack any usefulness from a Decision Support perspective. A third hope is to discourage the development of design- and /or condition-specific quality assessment measures – as opposed to the generic QE-5D - on the ground that they will almost always lack decision validity.

1. Quality weighting – the current position in meta-analysis for knowledge support

Randomised studies

The recent survey by Sutton et al. (Sutton et al., 2000) is a useful starting point because the Leicester group is also heavily involved in comprehensive decision modelling for decision support (see section 2). Beginning with interventional designs, they note that a plethora of quality assessment instruments have been constructed following the pioneering work of Chalmers. (Chalmers et al., 1981) Some of these produce an overall score, albeit without any explicit and theoretically-grounded weighting scheme, others simply a checklist/profile of component quality assessments. Of the clinical trial scales, Moher et al. regard only the Jadad one as 'developed according to accepted methodological principles' (Moher et al., 1999)and, perhaps not surprisingly, it has been shown that the various scales produce very different results when applied to the same collection of studies. (Juni et al., 1999) (Colle et al., 2002)

Sutton et al. join what they see as the growing support for concentration on *component* quality indicators to be used only in sensitivity analyses to gauge the influence of each individual quality adjustment on the unadjusted result. The authors' reasons for rejecting the *overall* index approach (which requires aggregation of the component assessments and hence weighting) are that:

> … (a) in each type of study, the factors influencing internal validity are likely to differ, and so standard scores will not be generically relevant; (b) the influence of factors affecting validity will differ, depending on the question and context of the trial; and (c) studies do not uniformly report sufficient details of the methods used in their design, conduct and analysis, to be able to accurately measure these factors in the same way in each study. (p134)

If, however, overall quality scores were to be derived for each study in a set, Sutton et al. accept that these could also be used by the meta-analyst to explore and report the effect of the study-specific quality adjustments on the overall conclusion. The important thing is that the quality adjustments would still not be incorporated into the conclusion (e.g. about the value of any particular parameter) by weighting.

Such incorporation (which, it may be recalled, is what we are advocating) is *possible* but undesirable according to Sutton et al.:

> Caution has been expressed in incorporating or using quality score in the weighting given to studies, because, while weighting study estimates by the precision has desirable statistical properties, quality scores are not direct estimates of precision, and *this approach lacks statistical or empirical justification* (emphasis supplied) (p140-1)

This final assertion seems now widely accepted, though others occasionally imply that 'weights' *could* somehow be 'validated' – usually without suggesting how. For example, Colle et al. in their recent replication of the Juni et al. study in the context of exercise therapy and low back pain, conclude that the 16 scales they included

> ‘usually lack a focused theoretical basis, and their objectives are unclear. They differ widely in the dimensions they cover and in their size and complexity; furthermore, the weight assigned to the key domains most relevant to the control of bias varies widely among them.(p1745) … For example, the Imperiale and McCullough scale score is dependent on inclusion and exclusion criteria (40%), whereas the Kleijnen scale score mostly reflects quality of randomization (20%), blinding of patients and care providers (20%), and number of patients included (30%). More generally, some scales include items that are related to the quality of presentation, ethical issues, and interpretation of results, whereas other scales contain items more related to the internal validity of trials. Whatever the type of scale, the scoring of items remains problematic.(p1749)

From our point of view Colle et al. are also interesting insofar as they suggest that different quality assessment scales should be used to assess pharmacologic interventions and physical therapies – and, by extension, suggest that a variety of condition- or intervention- specific quality scales are needed.

> Another unresolved question concerning the quality scale is whether 1 instrument can assess all situations. In other words, should pharmacologic intervention trials be assessed with the same instrument as physical therapy trials? Most of the quality scales used for meta-analysis or systematic reviews have been developed to assess drug therapy. This fact is reflected by the weight given to blindness of patient, of care provider, and of observer, which can account for up to 40% of the score. For physical treatments and particularly for exercise therapy, blindness of care provider or patients is usually impossible to obtain, as shown by the present systematic review: blindness of care provider was not achieved in the 20 trials, and blindness of patients was obtained only in 2 trials. Moreover, with patients knowing their treatment, it is difficult to ascertain the blindness of an observer…Assessment of physical therapy trials probably requires specific quality scales. (p1751)

On this question Colle et al. appear to be in line with a general tendency to favour design and/or condition specific quality scales – if one is going to have them at all. This is a tendency, it may again be recalled, we will also be questioning.

Non-randomised studies

Sutton et al. note that it is uncommon for interventional and observational studies to be combined in the same meta-analysis, essentially because the latter are regarded as more prone to various biases than the former. In a manner consistent with their concerns expressed above, they argue that, since no single quality scale can be relevant to all study designs, the weighting of studies using a quality score is particularly problematic if more than one study type is being combined.

In the context of epidemiological studies it is Greenland (Greenland, 1994a; Greenland, 1994b) who has shown the greatest hostility to 'quality weighting'.

> … quality scores are at best useless and at worst misleading (p301)… Perhaps the most insidious form of subjectivity masquerading as objectivity in meta-analysis is 'quality scoring' … [which] degrades data information by using arbitrary judgments to mix disparate items. … I wholeheartedly condemn quality scores because they conflate objective study characteristics (such as study design) with subjective and often arbitrary quality weighting schemes. Use of such scores can seriously obscure heterogeneity sources … and should be replaced by direct regression or stratification on objective quality-related study characteristics, such as study design (cohort, case-control, etc.), sources of data (direct interviews, mailed questionnaire, medical records, etc.) and sources of subjects (registry, hospital, etc.) (pp295-6) …with proper analysis of quality score *components*, quality scores … are superfluous and, without component analyses, quality scores can be misleading (p300)

This is all part of Greenland's condemnation of the 'synthetic' view of meta-analyses, in which they are seen as a method for producing a single estimate of effect from disparate study results. In his view a meta-analysis should be treated as a study of studies, rather than as a means of combining study results into a single effect estimate. This position is completely consonant with a Knowledge Support perspective, but we will see later that it is in direct conflict with the synthesising imperative of comprehensive decision modelling for Decision Support.

In their 1996 survey of research synthesis concentrating on observational studies in the field of public health, Mosteller and Colditz (Mosteller & Colditz, 1996) initially seem inclined to a decisional rather than knowledge perspective. They state that 'public health decisions are made on the available evidence … whether strong or weak…. Because evidence bearing on important public health issues is not always available from RCTs, we emphasise the "best evidence" approach to research synthesis' (p2). But their ultimate view is the conventional one:

> Although many have suggested using the quality of studies to adjust their reported outcomes or to weight them in meta-analysis, these ideas have not been given a strong statistical basis and are not currently being used in that way (p8) … Detsky et al. (Detsky et al., 1994) suggest that the field is not ready to provide such weights or adjustments… (p9)
>
> … results should not be combined across study design but summarized separately for each design... Quality may be used as a covariate to explain variation in study results. As of this writing, *no analytic procedure offers grounds for quality weighting.* (p19-20, emphasis supplied)

Shortly after Mosteller and Colditz reported the absence of any quality scoring system for observational studies paralleling the many developed for trials, Downs and Black proposed a quality scale that encompasses both interventional and observational studies. (Downs & Black, 1998) Like most of the RCT scales, this includes substantial weighting of items directed at the reporting as well as the substantive attributes of the study. The quality of reporting is, however, irrelevant from a decision support perspective, since the decision technologist is the person processing the study. But, in marked contrast to all previous quality scales, and definitely relevant from the decisional perspective, the Downs and Black scale embraces *external* as well as *internal* validity. Three of the 27 items are topic-specific, but the rest are generic to any health care study. An overall score is produced by summing the individual item scores and a global quality score is also elicited. (In their empirical application the two overall scores correlated well, but the global score was supplied by respondents after they had completed the item checklist so this was not unexpected.)

Downs and Black, in speaking of reviewers and readers 'being alerted to the particular weaknesses and strengths of a paper through the quality assessment' are clearly seeking improved Knowledge Support for a TIABIM DT . They do not contemplate quality weighting of the results using the overall scores, but do show considerable interest in the weighting issue:

> On the basis of current knowledge, we suggest assigning equal weighting to each of the five dimensions [Reporting, External Validity, Internal validity-bias, Internal validity – confounding (selection bias), and Power]. This is based more on the lack of evidence to prioritise one dimension over another rather than on any evidence to suggest each dimension was of the same importance. There are several ways forward. The simplest would be a sensitivity analysis in which the effect of adopting different weightings on the rating and ranking of studies could be observed. A more theoretical approach would entail some form of consensus development among experienced health care epidemiologists in which their *views of the relative importance* of the five dimensions were considered. Ultimately we need greater knowledge and understanding of the actual impact each dimension has on the effect size of the intervention being studied. (p381, emphasis supplied)

The key thing to note here is that quality assessment is seen as a technical matter of impact *estimation*, not one involving *preferences*. The suggestion that the *views of the relative importance* of experts might be sought is literally capable of either interpretation, but the intended interpretation is made pretty clear by the final sentence.

A more recent foray into quality assessment of non-randomised studies is the Newcastle-Ottawa Scale (NOS), an ongoing collaboration between the Universities of Newcastle, Australia and Ottawa, Canada. (Wells et al., 2003) There is no indication of any awareness that value judgments and preferences must underlie its 'star' weighting system and that differences in these will be a *legitimate* source of the inter-observer variation that worries their developers. In a recent presentation the group state that their key aim is to 'identify a threshold score distinguishing between "good" and "poor" quality studies' confirming that the mirage of providing an objective scoring method to distinguish high and low quality studies is alive and well.

It is present in the recently released Health Technology Assessment monograph by Deeks et al. Along with many other 'knowledge technologists' Deeks et al. recommend the development of *better* quality assessment tools, but show no recognition of the need for a *systematic* approach to their underlying basis in preferences. This is a little strange, because early on they cite Cooper (Cooper, 1984)as arguing that 'there are two main sources of variance in evaluator's decisions [on quality]: the relative importance they assign to different research characteristics and their judgments about how well a particular study meets a design criterion.' (Deeks et al., 2003, p23). Later they express the view that tools which follow 'the lines of Cooper's 'mixed-criteria' approach, requiring objective facts about a study's design followed by a quality judgment, may prove to be the most useful.'

Much of this debate in the context of health care and services has been paralleled in areas such as education, social care and work and criminology. The 'Threats to Validity' approach of Campbell and Cook has long been dominant and continues to be so. (Farrington, 2003; Shadish et al., 2002; Shadish & Haddock, 1994; Wortman, 1994). Space prevents detailed contact. Suffice to say that members of the Campbell Collaboration seem as unattracted by the idea of preference-based quality weighting as those of the Cochrane.

To sum up this section. Despite past controversy and some continuing disagreement the general consensus among 'knowledge technologists' seems in favour of assessing and reporting on the quality of individual study components for each study in a set, rather than producing an overall index quality score for each study. If the latter is produced, its role should be restricted to use as a covariate in sensitivity or regression analysis. This view is completely consonant with an interpretation of the task in hand as one of providing Knowledge Support, where the knowledge supplied to decision makers is that which has passed the criteria and standards appropriate to a Knowledge Technology. The tasks of 'taking into account' the evidential quality problems and 'bearing in mind' the results of the sensitivity analyses supplied as part of the Knowledge Support are left to be undertaken by decision makers using their decision making capacities, capacities that are *assumed* to be fully up to these tasks.

The only problem with this view - apart from the fact that it rules out the analytic DT that could be used if and when the final assumption does *not* hold - is that much of the discussion suggests a confusion between the estimation (*description)* of the types and magnitudes of bias present in any study and the *evaluation* of these within and across studies and, a fortiori, across study designs. From perusal of this literature it seems that many, perhaps most, knowledge technologists fail to have a sufficiently clear conceptual distinction between Knowledge Technologies, Valuation Technologies and Decision Technologies and that this is probably one of the major sources of continuing debate within their ranks over quality scoring and weighting. Two prevalent symptoms of the confusion between KTs and VTs are the failures to realise, or sufficiently emphasise, that both *exclusion criteria* and *multi-dimensional effect/outcome measures* necessarily require the use of 'subjective' value weights. Implying that these can be dealt with by adopting properly 'objective' KT methods, without use of 'subjective' VT methods as well, should not fool anyone. The failure to carefully distinguish the descriptive and evaluative components of 'quality weighting' is an equally serious problem. The impression is created that this could in principle be a value-free exercise, with an *objective* set of quality weights being a legitimate aim, even if one that would be extremely difficult to achieve.

Have those more interested in Decision Support come to a clearer and more coherent view on quality weighting?

2. Quality weighting – the current position in decision modelling for decision support

Sutton et al. freely admit that as a result of what happens in the conduct and analysis phases (as distinct from design phase) the results from a well-conducted and analysed cohort study may be more valid than those from a poorly conducted and analysed RCT. So, in deriving parameter estimates for a comprehensive decision model, how should the analyst deal with these two sources of evidence?

Despite their views reported earlier, which we interpret as in line with a Knowledge Support philosophy, Sutton et al. indicate a keen interest in the issue of quality weighting in comprehensive decision modelling. They survey the Confidence Profile Method, (Eddy et al., 1992) Cross Design Synthesis and Bayesian hierarchical models in this connection and we will follow this order.

Confidence Profile Method

The CPM provides a very general method for combining virtually any kind of evidence about a parameter and is therefore a key reference for our own argument, which is why we now look at what Eddy et al. have to say in detail - despite the fact that Sutton et al. conclude

> … uptake of the use of the method has been relatively poor, quite possibly due to its radically different conceptual approach to meta-analysis… A potential drawback of the method is that it is necessary to model biases in individual study estimates explicitly. Identification and quantification of such biases will usually be very difficult in practice. (p266)

Eddy et al. in the original work claim that the historical significance of the CPM lies in the way it offers a third alternative to analysts. (In the extracts that follow we italicise phrases that are crucial from the point of view of our later argument.)

> The notion of adjusting for biases, and the specific models for adjusting for biases, are new to the Confidence Profile Method. At present, there are no alternative quantitative methods that provide a comprehensive approach to this problem. The only alternatives are subjective. The usual approach to biases is ‘take it or leave it’; when an experiment is affected by one of more biases, the analyst can decide whether to include it or exclude it from the analysis. Another approach that has been applied to some meta-analyses is to ‘weight’ the evidence. *This is considered a subjective rather than a quantitative approach because there is rarely any theoretical basis for assigning the weights.* (emphasis supplied)
>
> The Confidence Profile Method enables analysts to develop models that adjust for biases, appropriate to the complexity and *perceived importance* of the biases that affect a particular problem…. (150, emphasis supplied)
>
> A method sometimes used to ‘adjust’ for biases is to assign a numerical ‘weight ‘ to the piece of evidence, with the intention of modifying its influence on the parameter to be estimated, relative to other pieces of evidence. The weight is usually registered as a number

> between zero and one, with one designating full weight (implying there are no biases) and zero designating no weight (implying the study is rendered valueless by biases). Weighting has the effect of decreasing the effective precision (e.g., enlarging the confidence interval) of a study without changing the point estimate. In effect, the question that must be answered by the person assigning the weight is, 'How does this collection of biases, taken together, affect the variance of the estimate of the magnitude of the effect?'
>
> Weighting suffers from two main problems. First, there is *rarely any theoretical or intuitive basis* for answering the type of question required to determine the appropriate weight to apply to a study. Second, when applied to a single experiment, weights can only increase the confidence limits (or variance of the distribution) for a parameter; they cannot adjust the maximum likelihood estimate (or mean of the posterior distribution). This is inaccurate because most biases affect the observed magnitude of a parameter. For these reasons the use of weights for bias reduction is strongly discouraged in the Confidence Profile Method.

The second reason is clearly irrelevant to 'cross-design synthesis' where the weighting is applied across studies, so we are reduced to considering the now familiar accusation that weighting lacks an acceptable basis. But this relates to the implicit confusion of statistical and value issues in the CPM. Immediately following on from the previous quote we read:

> With these caveats, there are functions that have indeed the effect of a weight. One of the simplest is to raise the likelihood function to a power, ...for example ½ would cause the evidence to have one half its unadjusted weight in the sense that if two identical pieces of evidence were combined each with a weight of ½, the combined likelihood would be equivalent to the unadjusted likelihood form one of the studies. (p150)

So *functional adjustments*, which have the effect of weighting, are permitted in CPM generating an overall impression of ambiguity in relation to subjective judgements. But to the extent that they are accepted their basis remains problematic. Why raise the likelihood function to a particular power? And what is to be the basis of the 'perceived importance' referred to in one of the quotes? No mention is made in CPM of the need for *value judgments* in respect of these subjective judgements, or generally in integrating different biases. There is no recognition of the fact that the overall conclusion may differ from person to person as a result of differences in *preferences* as well as differences in their assessments of the magnitude of each bias. We suggest that separating out the description and estimation of individual biases from their evaluative weighting is essential to clarification and progress.

Cross Design Synthesis

Drawing heavily on the work of Eddy et al. the US General Accounting Office has sought to develop a Cross Design Synthesis methodology. (US General Accounting Office, 1992) However, they appear to have got bogged down in the difficulty of carrying it out to the rigorous requirements of a KT – not surprisingly since in our view this is not just a difficult task but an impossible one. CDS accordingly remains a conceptual proposal rather than a practical method, amounting to little more than an injunction to identify and adjust for the biases characteristic of particular study designs (e.g. RCTs lack generalisability; data bases lack internal validity)

Bayesian Hierarchical Models

Given that they are making the case for Bayesian approaches it is a little strange to find Spiegelhalter et al. (Spiegelhalter et al., 2000)suggesting that the need in CPM to make explicit subjective judgments concerning the existence and extent of possible biases has perhaps restricted its application (p49). On the one hand, they seem fully supportive of the CPM's desire to model explicitly all the external and internal biases in each study, but on the other are clearly concerned about the assumptions necessarily required to establish the extent of the biases (p44), suggesting that these assumptions need some 'evidence base' in order to be convincing. So, while generally supportive of the Bayesian approach, there are frequent signs of an underlying Knowledge orientation in the form of worries that including studies with poorer designs will 'weaken an analysis' (p48). Of course it will - from a Knowledge perspective. Aside from isolated references to 'downweighting' of studies in the derivation of 'priors' (e.g. on the basis of their date) there is no sustained discussion of quality weighting in general, let alone any suggestion that any 'weights' require a basis in preferences.

However, Spiegelhalter and Best, (Spiegelhalter & Best, 2003) in a recently published paper on Bayesian approaches to multiple sources of evidence in complex modelling, do talk of generalised syntheses of available data 'in which multiple sources of evidence can be differentially weighted according to their assumed quality' (p3687). The degree of downweighting of potentially biased studies is, they say, in line with much we have heard before, 'a judgement that should be subject to sensitivity analysis' (p3688).

They provide an empirical example in which different quality weights are assigned to registry, RCT and case series studies within a random effects model of hazard ratios for alternative hip prostheses.

> As a baseline assumption for the quality weights we take… 0.5 [registry], 1.0 [RCT] and 0.2 [case series]… This corresponds to assuming that the 'bias' in the registry and case series studies leads to a 2-fold or 5-fold increase in the … variance, respectively over and above the between-study variability expected for RCTs (p3702).

An alternative set of weights, further downgrading the non-randomised data, increased the uncertainty (see table).

Registry weight	RCT weight	Case Series weight	Mean	CI
1	1	1	.54	(.37 - .78)
.5	1	.2	.61	(.36 - .98)
.1	1	.05	.82	(.36 – 1.67)

Where might these weights come from? Spiegelhalter and Best suggest

> Estimates or prior distributions of the between-study variances … and the quality weights … might be obtained from a possible combination of empirical random-effects analyses of RCTs of this intervention, historical 'similar' case studies, and judgement. (p3700)

Putting this alongside their assertion that the degree of downweighting of potentially biased studies is 'a judgement that should be subject to sensitivity analysis' we can see our two differences with Spiegelhalter and Best – and with the majority of other analysts in this area - neatly encapsulated. First, quality downweighting *is* a 'judgement' but it is a *value* judgement, a matter of preference. The use of 'judgement' without a qualifier is an indicator that beliefs and values are not being conceptually distinguished in the necessary way. Second, stressing the importance of sensitivity analysis in relation to quality weights in particular suggests that the fundamental commitment is still to Knowledge Support rather than Decision Support. If quality weights are based on theoretically-grounded, empirically-derived preferences we see no reason to believe sensitivity analysis is any more relevant in relation to quality adjustments than to any other aspect of the analysis. (Our personal view is actually that sensitivity analysis of the conventional sort has little or no place in a truly analytic decision technology when decisions have to be made in a limited time window. But that is not a topic for this paper)

The formal details of the Spiegelhalter and Best approach, along with the mathematics of our own proposals, including details concerning the treatment of uncertainty, are presented in a complementary paper. (Chalabi & Dowie, 2003)

Finally in relation to comprehensive decision modelling we look at the position taken by the Leicester group. Cooper et al. (Cooper et al., 2003b) outline the way in which evidence from a variety of study designs and data sources can be drawn on in comprehensive decision modelling.

> … comprehensive decision analytical economic models (i.e. integrating the synthesis and decision process together into a single coherent model) may be implemented within a Bayesian framework (pX1) … Desirably, the methods … should be extended to … models which incorporate data from different study designs (e.g. RCTs, observational studies together with *expert judgment as to the relative merit of different sources of evidence*) … (pX19; emphasis supplied)

They cite approvingly the work by Parmigiani and colleagues on the Stroke Prevention Policy Model where

> … to inform the natural history part of the model data is extracted from major epidemiological studies, whilst to inform the effectiveness of different interventions data is based on a literature review, meta-analyses of trials, administrative insurance claim records and patient utility data. (pX4)

However, the approach commended by Parmigiani and the one apparently used in Leicester's own work, (Cooper et al., 2003a; Cooper et al., 2002) is essentially a lexicographic one, in which the 'best available evidence' is defined as that available at the highest possible level of the evidence hierarchy, not that which results from according evidence from all levels appropriate weight.

In relation to the justification of the introduction of expert judgment as to 'the relative merit of different sources of evidence' in forming priors we can note that 'merit' is a preference-based concept. Elsewhere two members of the group (Sutton & Abrams, 2001) provide a further demonstration of the way in which preferences (subjective values) are implicitly (unconsciously? unwittingly?) transformed into priors (subjective beliefs):

> The observational evidence can be discounted … and given less weighting than the randomized trials. Such a situation would be appropriate if a researcher believes that although the observational evidence provides some information …, concern that serious biases may exist means that it should be treated with caution. Such a prior is labelled 'sceptical' since caution is being expressed due to this down-weighting. For [this] example… a variance four times larger than that of the randomised trials is used… (p294)

Caution is clearly a preference concept - a 'risk preference' – in the same way as *merit*.

Sutton et al. point out that the reliance on investigator *judgments* is a major limitation of the CPM method. However, things are different when they turn their attention to Bayesian Hierarchical Models, where we read

> … the Bayesian approach can accommodate a priori *beliefs* regarding qualitative differences between the various sources of evidence. These prior distributions may represent subjective *beliefs* elicited from experts, or other data-based evidence. …*beliefs* about the relative merits of individual studies or types of study can be incorporated in the model. For example, *beliefs* about the relative value of RCTs, cohort-study and case-control study results may be modelled explicitly, and the dependence of the conclusions of the review on these *beliefs* investigated (p268; emphases supplied)

This is the crunch. Treating quality assessments as *belief* differences to be explored in sensitivity analysis of different priors leaves us in a Knowledge Technology/ Knowledge Support framework, albeit a less rigorous one embracing a much wider hierarchy of evidence. Our view is that we need to treat these quality assessments as matters of *preference* in order to produce an answer (decision support), not just a series of alternative possibilities. (Of course, while requested and interpreted as different *beliefs* the different priors elicited from a group of experts probably reflect their different *preferences* as well!)

Perhaps reflecting their disciplinary origins in statistics rather than economics, the 'comprehensive decision modellers' have not – at least so far - gone beyond the 'knowledge technologists' in relation to quality weighting.

3. Quality weighting of parameter inputs into a Bayesian Decision Technology

The alternative approach that we are canvassing is one that comes into play after any feasible and simple *statistical* adjustments for different types of bias have been made – in the (highly unlikely) extreme situation, after all the adjustments envisaged by the CPM method and later advances (Ades, 2003) have been undertaken. This alternative approach is an explicitly *preference-based* one. In it the judgements necessary to pool the parameter estimates from different sources are *value judgements,* value judgements concerning the relative disutility of the different types and magnitudes of deviations from the relevant ideal, non-biased, study that are exhibited in the evidential sources actually available *within the decision window.*

Explicit subjective belief (descriptive) judgments about the existence and extent of possible biases need to be carefully distinguished from explicit subjective value judgments about the relative 'quality' weights to be attached to varying types and extents of bias. A subjective belief judgment about whether a study shows a small or large amount of selection bias or a small or large amount of attrition bias is quite a different matter from the relative 'disutility' assigned to a small amount of selection bias compared with a large amount of attrition bias. The assignment of 'quality weights' requires both description and evaluation. This critically important conceptual distinction is missed and/or implicitly denied in all the literature we have surveyed to this point.

Within a Decision Technology the need for subjective judgments of *both* sorts represents the normal condition of life, rather than the fundamental 'problem' it poses for a Knowledge Technology. This is not to say that the 'hierarchy of evidence' is not to be respected in general, simply to emphasise that the use of quality *cut-offs* rather than quality *weights*, and the implicit denial that value judgments are necessarily involved in any multi-dimensional aggregation, represent futile attempts to adhere to strictly statistical criteria for quality assessment. There is no need to resort to 'arbitrary' weights as the alternative, because there are well-established methods – valuation technologies - for eliciting preferences of all kinds.

We take it as virtually axiomatic that parameter derivation in a comprehensive Bayesian decision model should be based on quality weighting of all relevant evidential sources. The absolute weights run from 0 to 1, though only the relative weights are relevant in the pooling process for each parameter (see table 3). Once quality weighting has been accepted we can see no justification for either a lexicographic or a threshold approach. Both involve giving zero weight to studies which *by definition of the quality scale* should be accorded some weight. (This is not, of course, inconsistent with some studies being rated at zero on the scale).

Assuming that stochastic Bayesian modelling is being employed as a *Decision* Technology – a way of identifying the optimal choice among competing options in some defined decisional time window – the central task is to arrive at the *best available* estimates for each parameter (or parameter distribution), irrespective of the state of the evidence according to *Knowledge* Technology standards. Establishing the best available estimate may require descending below any 'statistically acceptable' cut-off and the complete 'hierarchy of evidence', from RCT to opinion, is potentially relevant. Evidence relevant to a parameter is likely to reside in multiple sources covering a variety of study designs and data resources. Each will be subject to its own known general strengths and weaknesses ('internal' biases), as well as ones *particular* to the study or data base concerned ('external' biases relative to the precise population/condition being modelled, as well as 'internal' biases).

Following from the arguments of the previous sections we suggest (i) that the best available parameter is that which accords appropriate (i.e. quality-adjusted) weight to all available evidence on that parameter, and (ii) that the necessary quality weighting of evidence should be considered to be a matter of *generic preferences* over the possible ‘evidential states’. We offer a simple instrument (QE-5D) that will enable a (linear) ‘tariff’ or series of tariffs to be developed for a comprehensive set of evidential states.

An analogy will be helpful.

Quality weighting of health states is undertaken by way of multidimensional instruments that eventually locate each state on a scale running from 1 (best possible state, often ‘full health) to 0 (worst possible state, often ‘dead’). For example, the EQ-5D instrument has 5 dimensions (mobility, self care, usual activities, pain and discomfort, anxiety and depression) and 3 levels on each (1, no problems; 2, some problems; 3, severe problems). This generates 243 health states, for each of which an evaluation exercise yields a tariff value for Health-Related Quality of Life - the ‘Quality Adjustment’ element of a QALY calculation. For example, the EQ-5D state comprising some problems on all 5 dimensions, coded 22222, has a York tariff of 0.516.

By analogy QE-5D will have five dimensions (external bias, selection bias, performance bias, attrition bias and detection bias) and 3 levels on each (1, none; 2, a small amount; 3, a large amount). This will generate 243 ‘evidential’ states, for each of which an evaluation exercise/survey will yield a tariff value for ‘Knowledge-Related Quality of Evidence’. The extreme states are 11111, presumably the results of a perfect RCT that one would eventually locate at 1 on the scale, and 33333, the results of a ‘study’ (or parameter source) with large amounts of all five sorts of bias, that would eventually be located at 0. (Incidentally, we can think of no conceptual equivalent of ‘Dead’ and hence of ‘States worse than Dead’. If the worst possible state (pole) on the KRQOE scale is ‘worthless’ it is hard see how one could have evidence which is ‘worse than worthless’, or, if one could, how one could give it negative weight in parameterisation.)

This tariff will provide absolute quality ratings, but for any particular parameterisation it will be the relative weights of the available sources that determine the result. It will be irrelevant whether, for example, given that two sources are available, they are rated .2 and .1 on the tariff or 1.0 and 0.5. In deriving the central value of the parameter they would be weighted 2/3 and 1/3 respectively.

The descriptive form of QE-5D appears as Table 1.

The five *dimensions* are made up of the four internal biases as classified in the Cochrane Handbook, to which is added external bias. They appear to cover the major components of all other quality scales. Others may wish to argue for different dimensions and, subject to the acceptance that we are in a Decision Support and DT context, we regard this debate as open. However, if, as in the proposals of Farrington, we are offered a five dimensional scale comprising the 4 Shadish/Campbell-Cook dimensions plus a reporting dimension, (Farrington, 2003) it is important to note that dimensions other than internal and external validity are relevant only in a Knowledge Support and KT context.

The three qualitative *levels* of QE-5D are merely an initial proposal and are also open for discussion and revision. It is clearly vital that the levels cover the entire range of the evidential possibilities and that the second lowest level is defined in such a way as to permit some positive value to 'expert opinion'. In other words one cannot have levels that mean expert opinion/judgment will always be assigned to the lowest level on any dimension. In order to achieve some discrimination within randomised and non-randomised studies it may therefore be necessary to use 4 or even 5 levels. The problem with such expansion is that the 243 cells in the 3 level matrix (given 5 dimensions) become 1024 cells with 4 levels and 3125 with 5. So, while, as with HRQOL instruments, there will doubtless be calls for more dimensions and more levels, we feel that, as there, the trade-off between practicality and complexity has to favour the former. That is why we initially propose just three.

Most of the existing scales simply contain too many dimensions and/or levels to be practical from the preference elicitation point of view, as well as being largely design-specific. One is also entitled to entertain some doubts as to whether pursuing the additional complexity they involve - entirely appropriate on KT grounds - will have substantive consequences for decisional evaluations, where the task is merely to establish the best option.

As in the HRQOL field there will be calls for 'design-specific', 'intervention-specific' and 'topic-specific' instruments and we have already seen hints of these in quotes in earlier sections. Colle et al., for instance, argue that it will be inappropriate to use the same scale for pharmacotherapy and physical therapy studies. Some quality components (e.g. blinding) will not, and indeed cannot, achieved high or possibly any score in the latter. Therefore, the conclude, the quality scores will be 'biased' if the same scale is used for both. We believe this to be a faulty conclusion. For various reasons, evidence from studies of some types of intervention may indeed be incapable in principle of achieving the same absolute KRQOE score as others . However, this is actually a reason for insisting on *generic quality* weight scales for all parameters in all arms of all decision models. As in HRQOL, decisional criteria not knowledge criteria should determine the instrument used. (Dowie, 2002) Wherever a decision model involves any parameter that requires a generic instrument - because the evidential sources for it cover the full evidential hierarchy - methodological coherence requires that the generic instrument should be used for *all* parameters. It seems highly unlikely, especially when utilities and costs are part of the analysis, that this will not be the case in any decision model.

The *evaluative* form of QE-5D would be administered to populations of Cochranites, Campbellites, Health Economists and other relevant professional subgroups, rather than to the general population, though (in contrast to the suggestion of Downs and Black) this would be because of the practical need of respondents to *understand* the dimensions and levels of the instrument, not because these professionals' utilities are conceptually superior or more appropriate.

A limited set of 15 evidential states would be administered to each individual respondent, using criteria not dissimilar to those used by the EuroQol developers. (Williams, 1995, pp17-18)

Which Valuation Technology should be used? Visual Analog Scale and Willingness to Pay VTs are relatively easy to envisage.

It is also possible to conceptualise a Standard Gamble VT in the evidential context, along the following lines. We need a central estimate for a parameter. The rate given by evidential state X (based on an extensive case control series) is already available to the person/s whose QE tariff you regard as relevant. They are told that in response to a request sent to a website medical Q and A service, you receive a reply which provides either the results from a high quality RCT or the personal opinion of the unidentified person who handled the query. Unfortunately it is not clear from the message which of these it is. Assuming time or other reasons prevent you clarifying the message, and assuming that you would regard the personal opinion of the unidentified person as the worst possible evidence, how certain would you need to be that the rate in the message was that from the high quality RCT in order to use it instead of the rate of known provenance you already have. Your minimum probability will indicate your QE rating for evidential state X.

Equivalent Time Trade -Off and Person Trade-Off approaches are not immediately obvious.

Whichever VT is used it is vital to see that no mention will be made in this instrument of study design characteristics, i.e. terms such as randomised or non-randomised will never appear. We are seeking evaluations of extents and types of *bias*, not of study designs (or study conduct or study analysis). It may be expected, of course, that randomised studies will usually emerge with a higher QE-5D rating than non-randomised ones. But it may well be that an RCT with a large amount of external bias may be out rated (in the tariff) by a case control study with no external bias but greater amounts of internal ones. In advance of the preference elicitation study we don't know the answer. Different tariffs will, as in assessments of HRQOL, produce different conclusions.

Conclusion

We will be interested to hear if 'the field' is any more ready to accept empirically grounded, preference-based weighting in 2004 than it was to accept 'arbitrary' belief-based weighting when Detsky et al. wrote in 1994. If we take the emanations from the Cochrane and Campbell Collaborations to represent 'the field', it seems unlikely. The current lack of awareness of the necessarily *preferential* origins of quality weight indices based on multiple dimensions and levels – and the consequential need for the introduction of explicit preference elicitation using valuation technologies - is staggering.

In his insightful review of quality weighting in the criminology area Farrington suggests that it is important to develop some simple kind of

> 'index of methodological quality in order to communicate to scholars, policy makers and practitioners that not all research is of the same quality…It seems highly desirable for funding agencies, journal editors, scholarly associations, and/or the Campbell Collaboration to get together to agree on a measure of methodological quality that should be used in systematic reviews and meta-analyses in criminology. (pxx)

Setting aside our disagreement with the lumping together of knowledge and decision-focused groups, the difficulties of getting all of the groups – or even all of those within any one of these groups - to accept that quality weights necessarily involve *value* judgments will undoubtedly be profound, given that at present the problem is either unrecognised, denied or diminished. Referring to his own proposed scale with 5 dimensions - internal validity, descriptive validity (reporting), statistical conclusion validity, construct validity and external validity – and 4 levels – very poor, poor, adequate, good – Farrington follows in the long tradition of those who have developed instruments producing an overall quality measure. He notes, almost in passing

> There are many ways of producing a summary score (0-100) from the individual (0-4) scale scores. For example, consistent with *my ordering of the importance* of the five types of validity, internal validity could be multiplied by 8 (maximum 32), descriptive validity by 6 (maximum 24), statistical conclusion validity by 4 (maximum 16), construct validity by 4 (maximum 16) and external validity by 3 (maximum 12). (pxx; emphasis supplied)

This lack of serious and sustained methodological interest in the issue of theoretically and empirically grounded 'importance ordering and weighting', arising either from a lack of relevant disciplinary background, or from a knowledge rather than decision focus - or from both - needs to be remedied.

Table 1: Descriptive form of QE-5D (3L)

EXTERNAL BIAS	
1 None	
2 A small amount	
3 A large amount	
SELECTION BIAS	
1 None	
2 A small amount	
3 A large amount	
PERFORMANCE BIAS	
1 None	
2 A small amount	
3 A large amount	
DETECTION BIAS	
1 None	
2 A small amount	
3 A large amount	
ATTRITION BIAS	
1 None	
2 A small amount 3 A large amount	

Table 2: Two evidential states from the evaluative instrument:

Study xx (code **22222**)		Study yy (code **21321**)
A small amount of **external** bias		A small amount of **external** bias
A small amount of **selection** bias		No **selection** bias
A small amount of **performance** bias		A large amount of **performance** bias
A small amount of **attrition** bias		A small amount of **attrition** bias
A small amount of **detection** bias		No **detection** bias

Table 3 Example of parameter derivation for deterministic model

Study	Parameter value	QE-5D code	London tariff	Relative weight	Parameter Weight
Bloggs 2003	.7	32332	.2	.06	.04
Cloggs 1997	.6	12111	.8	.27	.16
Floggs 2000	.4	22122	.6	.20	.08
Ploggs 1988	.8	22222	.5	.17	.14
Sloggs 1994	.5	11121	.9	.30	.15
			Normalization constant = 3	Sum of relative weights = 1	Pooled parameter value =**.57**

References

Ades, A. (2003), 'A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence', *Statistics in Medicine*, 22, 2995-3016.

Chalabi, Z. and Dowie, J. (2003), 'The mathematics of quality weighting in evidential synthesis', *in preparation*.

Chalmers, T., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. and Ambroz, A. (1981), 'A method for assessing the quality of a randomized control trial', *Controlled Clinical Trials*, 2, 31-49.

Colle, F., Rannou, F., Revel, M., Fermanian, J. and Poiredeau, S. (2002), 'Impact of quality scales on levels of evidence inferred from a Systematic Review of exercise therapy and low back pain', *Archives of Physical Medicine and Rehabilitation*, 83, 1745-1752.

Cooper, H. (1984), *The integrative research review: a systematic approach*, Newbury Park, CA, Sage.

Cooper, N., Abrams, K., Sutton, A., Turner, D. and Lambert, P. (2003a), 'Use of Bayesian methods for Markov modelling in cost-effectiveness analysis: An application to taxane use in advanced breast cancer', *Journal of the Royal Statistical Society Series A*, 166, 3.

Cooper, N., Sutton, A. and Abrams, K. (2002), 'Decision analytical economic modeling within a Bayesian framework: Application to prophylactic antibiotics use for caesarean section', *Statistical Methods in Medical Research*, 11, 491-512.

Cooper, N.J., Sutton, A.J., Abrams, K.R., Turner, D. and Wailoo, A. (2003b), 'Comprehensive decision analytical modelling in economic evaluation: a Bayesian approach', *Health Economics*, in press.

Deeks, J., Dinnes, J., D'Amico, R., Sowden, A., Sakarovitch, C., Song, F., Pettigrew, M. and Altman, D. (2003), 'Evaluating non-randomised intervention studies', *Health Technology Assessment*, 7, 27.

Detsky, A., Naylor, C.D., O'Rourke, K., McGeer, A.J. and L'Abbe, K.A. (1994), 'Incorporating variations in the quality of individual randomized trials into meta-analysis', *Journal of Clinical Epidemiology*, 45, 3, 255-265.

Dowie, J. (2001), 'Decision analysis and the evaluation of decision technologies', *Quality in Health Care*, 10, 1, 1-2.

Dowie, J. (2002), 'Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions', *Health Economics*, 11, 1-8 and 21-22.

Downs, S.H. and Black, N. (1998), 'The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions', *Journal of Epidemiology and Community Health*, 52, 377-384.

Eddy, D.M., Hasselblad, V. and Schacter, R. (1992), *Meta-Analysis by the Confidence Profile Method: the Statistical Synthesis of Evidence*, Boston, Academic Press, Inc.

Farrington, D. (2003), 'Methodological quality standards for evaluation research', *Annals of the American Academy of Political and Social Science*, 587, 49-68.

Greenland, S. (1994a), 'A critical look at some popular meta-analytic methods', *American Journal of Epidemiology*, 140, 290-296.

Greenland, S. (1994b), 'Quality scores are useless and potentially misleading', *American Journal of Epidemiology*, 140, 300-301.

Juni, P., Witschi, A., Bloch, R. and Egger, M. (1999), 'The hazards of scoring the quality of clinical trials for meta-analysis', *Journal of the American Medical Association*, 282, 1054-1060.

Moher, D., Cook, D., Jadad, A., Tugwell, P., Moher, M., Jones, A., Pham, B. and Klassen, T. (1999), 'Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses', *Health Technology Assessment*, 3, 12.

Mosteller, F. and Colditz, G.A. (1996), 'Understanding research synthesis (Meta-analysis)', *Annual Review of Public Health*, 17, 1-23.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston, Houhgton-Mifflin.

Shadish, W.R. and Haddock, C.K. (1994), 'Combining estimates of effect size', *in* Cooper, H. and Hedges, L.V. (eds.), *The Handbook of Research Synthesis*, New York, Russell Sage Foundation.

Spiegelhalter, D., Myles, J., Jones, D. and Abrams, K. (2000), 'Bayesian methods in health technology assessment', *Health Technology Assessment*, 4, 38.

Spiegelhalter, D.J. and Best, N.G. (2003), 'Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effective modelling', *Statistics in Medicine*, 22, 3687-3709.

Sutton, A.J. and Abrams, K.R. (2001), 'Bayesian methods in meta-analysis and evidence synthesis', *Statistical Methods in Medical Research*, 10, 277-303.

Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A. and Song, F. (2000), *Methods for Meta-Analysis in Medical Research*, Chichester, John Wiley and Sons.

US General Accounting Office (1992), *Cross Design Synthesis: A New Strategy for Medical Effectiveness Research*, US General Accounting Office.

Wells, G., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M. and Tugwell, P. (2003), 'The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses', *http://www.lri.ca/programs/ceu/oxford.htm*.

Williams, A. (1995), 'The measurement and valuation of health: a chronicle', *University of York Centre for Health Economics: Discussion Paper 136*.

Wortman, P.M. (1994), 'Judging research quality', *in* Cooper, H. and Hedges, L.V. (eds.), *The Handbook of Research Synthesis*, New York, Russell Sage Foundation.