Value judgments in the analysis and synthesis of evidence

Daniel Strech\textsuperscript{a, *}, Jon Tilbury\textsuperscript{b}

\textsuperscript{a}Institute of Medical Ethics, University of Tübingen, Germany
\textsuperscript{b}Division of General Internal Medicine and the Bioethics Research Program, Mayo Clinic, Rochester, MN, USA

Accepted 2 January 2008

Abstract

**Objective:** To describe the principal role of value judgments in the analysis and synthesis of evidence as they are involved in systematic reviews, meta-analyses, and health technology assessments.

**Method:** Using the tools of conceptual analysis, we characterize three main types of value judgments and propose an outline of how to enhance the appropriate role of value judgments in the process of analyzing and synthesizing evidence.

**Results:** The production, analysis, and synthesis of evidence involve value judgments characterized as preferences of persons or groups that cannot be validated by appeal to facts alone. Because preferences across individuals can vary, value judgments can be a source of bias in science and unwarranted variation in the application of scientific evidence. However, it is not possible or desirable to eliminate all value judgments in the process from production to synthesis of evidence.

**Conclusion:** With respect to the value judgments that shape the analysis and synthesis of evidence, review authors should disclose and justify choices related to the three key value judgments outlined in this paper. Authors should also highlight how their value judgments differ from the stated or implicit value judgments of previously published reviews on the same topic.

Keywords: Bioethics; Value judgments; Evidence-based medicine; Bias; Transparency; Systemic reviews

1. Evidence and value judgments

The production, analysis, and synthesis of evidence are scientific endeavors. Science seeks a better understanding of truth. In clinical biomedical science, this quest results in attempts to make accurate estimates of risk and benefit of health interventions to better approximate truths about health. In seeking the truth, scientists must take measures to minimize bias. The tools of biomedical research and evidence-based medicine help both designers and users of the evidence to minimize exposure to biases [1–4]. There also are well-described biases in the interpretation and communication of evidence [5,6].

Scientific evidence supports the search for accurate descriptions of what is the case. However, for science to influence decision making, descriptive “is” statements be combined with value judgments (VJs). VJs are preferences of persons or groups that cannot be validated by appeal to facts alone. They are a separate category of human reasoning that complements data from observation. Because preferences across individuals can vary, VJs can be a source of bias in science and unwarranted variation in the application of scientific evidence. However, all VJs in science do not necessarily result in bias. And although producers and interpreters of biomedical science should strive to eliminate bias, it is simply impossible and undesirable to eliminate all VJs in the production, analysis, synthesis, and application of evidence. This has already been recognized in different steps of the production of evidence such as determining adequate alpha errors or sample sizes, assessing content validity for quality-of-life scales, or incorporating patient preferences into randomized trials [7–10]. Other authors have called attention to the importance of VJs in the application of evidence. For instance, determining the levels of significance or thresholds for cost-effectiveness requires VJs [10–12]. However, there is still little awareness of VJs in the analysis and synthesis of evidence such as occurs in systematic reviews. In a recent interview with Sean Tunis, David Eddy argues that scientific judgments and VJs are rather separate [13]. He states that in the analysis of evidence we use scientific judgments, whereas later on, when applying the evidence to individuals or health policy decisions, VJs come into play. In contrast, here we contend that along the complete spectrum from producing, analyzing, synthesizing, and applying evidence, investigators face
VJs. If so, those VJs deserve explicit examination. Unlike earlier work we focus more specifically on VJs that occur in the process of evidence analysis and synthesis as occurs commonly in systematic reviews, meta-analyses, and health technology assessments. In so doing, we hope to further promote the dialogue about the role of VJs in clinical medical sciences.

As we outline in Table 1, VJs come into play in most steps along the spectrum of evidence production, analysis, synthesis, and application. The GRADE approach and the CONSORT statement give a more detailed explanation of these steps [3,14]. Whenever scientists use words like “appropriate,” “sufficient,” “significant” in presenting or appraising evidence, they are making VJs. Which level of significance is adequate? Which criteria are appropriate to judge the study quality? Are the inconsistencies within the study or among studies important? Should the study be excluded from meta-analysis? Despite their importance, the role of VJs particularly in the analysis and synthesis of evidence has not been discussed in detail.

Below we propose a categorization for three main VJs in the analysis and synthesis of evidence. Table 2 presents three basic types of VJs (Table 2). These general categories may not capture all VJs that come into play in the analysis and synthesis of evidence, but they do begin to clarify an important and underappreciated dimension of these processes.

2. The framework of value judgments in the analysis and synthesis of evidence

We can distinguish between three principal types of VJs in the analysis and synthesis of evidence: judgments about (1) choosing outcome measures, (2) balancing benefits and harms, and (3) tolerating uncertainty.

Table 1
Judgments in the production of evidence and their normative characteristics

<table>
<thead>
<tr>
<th>Judgments in the production, analysis, and synthesis of evidence</th>
<th>Examples (normative characteristics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptualization and simplification of patient relevant outcomes</td>
<td>What are the relevant outcomes?</td>
</tr>
<tr>
<td>Prioritization of research question</td>
<td>What degree of simplification in outcome measurement is warranted?</td>
</tr>
<tr>
<td>Study design</td>
<td>Which research question is worth studying and should be financed?</td>
</tr>
<tr>
<td>Importance of outcomes</td>
<td>Which study design is appropriate for the specific clinical question?</td>
</tr>
<tr>
<td>Analysis</td>
<td>Which outcomes (for benefits and harms) are more and which are less important?</td>
</tr>
<tr>
<td>Reporting</td>
<td>Which level of significance is adequate to the research question?</td>
</tr>
<tr>
<td>Study quality (selection and specification of assessment criteria and consistency in data)</td>
<td>Which degree of simplification within the reporting of study findings is appropriate?</td>
</tr>
<tr>
<td>Data synthesis (inclusion and exclusion of data)</td>
<td>Which findings should be explicitly reported and which can be ignored? Which models of reporting the effect size are appropriate?</td>
</tr>
<tr>
<td>Balancing benefits and harms</td>
<td>Which criteria are appropriate to judge the study quality?</td>
</tr>
<tr>
<td>Strength of recommendations (overall quality, downgrading)</td>
<td>Are the inconsistencies within the study or in relation to other studies important?</td>
</tr>
<tr>
<td></td>
<td>Which limitations/shortcomings in the study quality are sufficient for excluding or discounting the results?</td>
</tr>
<tr>
<td></td>
<td>Is the method for data aggregation appropriate?</td>
</tr>
<tr>
<td></td>
<td>Should the study be excluded of meta-analysis?</td>
</tr>
<tr>
<td></td>
<td>Are the benefits of the intervention are more important than the harms and vice versa?</td>
</tr>
<tr>
<td></td>
<td>What are the adequate reference points (outcomes for benefits and harms) to judge the overall quality? What framework for the “level of evidence” and “strength of recommendation” is appropriate?</td>
</tr>
</tbody>
</table>
The findings of clinical trials as well as systematic reviews always contain some degree of uncertainty due to biasing influences of chance and different types of biases. Accordingly, a third type of VJ that dominates the process of planning a clinical trial also influences the assessment of study quality in systematic reviews—balancing the value of gaining more but less accurate evidence vs. gaining less evidence that is more accurate. VJs are also important in choosing which research designs to take seriously. For example, choosing an experimental versus an observational approach, or choosing important effect modifiers (e.g., ethnicity or social class). Because even the most rigorous study designs cannot eliminate all uncertainty, those evaluating data from individual studies must decide how much uncertainty or potential for bias they are willing to accept when judging the merits of published evidence. These choices are not merely technical in nature but reflect the values of those choosing. Although it is common for investigators to use validated quality assessment instruments, there are no agreed criteria for specifying the weight that should be assigned to limitations in studies’ quality once assessed. Unless there are severe concerns, most systematic reviews include all studies in the sensitivity analyses to assess the possible influence of the various quality items [19]. Nevertheless, deciding whether deficits in measured study quality are important enough to exclude the study findings from meta-analysis, for instance, often reflects review authors’ tolerance of uncertainty.

Although the often-applied Jadad criteria focus on the randomization process, other review authors prefer criteria that focus on the degree of postrandomization exclusions or baseline imbalances [20]. Recently, for instance, Cochrane reviews of cholinesterase inhibitors in the treatment of patients with Alzheimer’s disease concluded an overall benefit for cognitive and global outcome measures after employing the Jadad score for quality appraisal of clinical trials [21–23]. Another systematic review came to a rather pessimistic conclusion after focusing on postrandomization exclusions, baseline imbalances, and a more critical analysis of the reporting of patient randomization [24].

These VJs are similar to the determination of alpha or beta errors in designing studies [10,25]. They can be specified a priori, but the exact levels chosen are determined by social convention or personal preference. We must acknowledge that every choice in this regard requires balancing the uncertainty of being wrong in our inferences about study quality with the probability of missing important signals about true benefits and harms from studies of suboptimal quality. The answer to how much uncertainty in study quality we are willing to accept ought to be dependent on the context (e.g., severity of disease, existence of alternatives) and on the preferences and values of the particular patient population to which the evidence will be applied. These VJs, therefore, have to be informed by knowledge in

---

### 2.2. Balancing benefits and harms

When investigators synthesize data into systematic reviews and guidelines they attempt to answer the question, “Does the intervention (as examined across multiple studies) do more good than harm?” The task of balancing benefits and harms is further complicated by the need to both accurately ascertain the magnitude of effect for a given outcome and estimate the quality of those research results for purposes of weighing. Balancing multiple dimensions of benefit and harm data—relevant outcomes, magnitude of effect, and quality of results—presents a particular challenge in formulating guideline recommendations that often depend on quality evidence synthesis.

Take the example of mammography screening [16]. After lengthy, often politically charged deliberations, the U.S. Preventive Services Task Force concluded that there was an overall benefit to mammography as evidenced by decreased breast-cancer specific mortality despite the harms of false positive test results [17]. On the other hand, a Cochrane review came to the opposite conclusion, citing no decrease in overall mortality after mammography and noting the harms related to increasing surgical interventions [18]. The differences between the U.S. Preventive Services Task Force and the Cochrane review conclusions illustrate differing preferences for certain outcome measures (i.e., VJs) leading to conflicting weighting of benefits and harms of mammography.

---

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value judgments and evidence synthesis</td>
</tr>
<tr>
<td>Three key value judgments</td>
</tr>
<tr>
<td>Choosing outcome measures—Which outcome measures are more or less important for the patient and for the evaluation of a certain intervention?</td>
</tr>
<tr>
<td>Balancing benefits and harms—Given the magnitude of effect for a given outcome and the quality of those research results, does the intervention do more good than harm?</td>
</tr>
<tr>
<td>Tolerating uncertainty—Deciding the merits of gaining more but uncertain knowledge vs. gaining less knowledge with greater certainty</td>
</tr>
</tbody>
</table>

---

"Does the intervention (as examined across multiple studies) do more good than harm?" The task of balancing benefits and harms is further complicated by the need to both accurately ascertain the magnitude of effect for a given outcome and estimate the quality of those research results for purposes of weighing. Balancing multiple dimensions of benefit and harm data—relevant outcomes, magnitude of effect, and quality of results—presents a particular challenge in formulating guideline recommendations that often depend on quality evidence synthesis.

− investigator places on each scientific and practical factor under consideration. Evaluating the merits of different outcome measures occurs at several points along the spectrum from evidence production to evidence synthesis. For instance, when designing a clinical trial, an analysis plan specifies in advance what the primary and secondary outcomes will be.

− These choices are then amplified and reflected when evidence from that research is eventually synthesized. In systematic reviews at the other end of the spectrum, the authors have to define their outcomes of interest and accordingly rate the merits of outcomes reported in studies as part of assessing inclusion and exclusion criteria. Preferences concerning suitable outcomes, therefore, have an impact on clinical trial designs and inclusion criteria for systematic reviews.

### 2.3. Tolerating uncertainty

The findings of clinical trials as well as systematic reviews always contain some degree of uncertainty due to biasing influences of chance and different types of biases. Accordingly, a third type of VJ that dominates the process of planning a clinical trial also influences the assessment of study quality in systematic reviews—balancing the value of gaining more but less accurate evidence vs. gaining less evidence that is more accurate. VJs are also important in choosing which research designs to take seriously. For example, choosing an experimental versus an observational approach, or choosing important effect modifiers (e.g., ethnicity or social class). Because even the most rigorous study designs cannot eliminate all uncertainty, those evaluating data from individual studies must decide how much uncertainty or potential for bias they are willing to accept when judging the merits of published evidence. These choices are not merely technical in nature but reflect the values of those choosing. Although it is common for investigators to use validated quality assessment instruments, there are no agreed criteria for specifying the weight that should be assigned to limitations in studies’ quality once assessed. Unless there are severe concerns, most systematic reviews include all studies in the sensitivity analyses to assess the possible influence of the various quality items [19]. Nevertheless, deciding whether deficits in measured study quality are important enough to exclude the study findings from meta-analysis, for instance, often reflects review authors’ tolerance of uncertainty.

Although the often-applied Jadad criteria focus on the randomization process, other review authors prefer criteria that focus on the degree of postrandomization exclusions or baseline imbalances [20]. Recently, for instance, Cochrane reviews of cholinesterase inhibitors in the treatment of patients with Alzheimer’s disease concluded an overall benefit for cognitive and global outcome measures after employing the Jadad score for quality appraisal of clinical trials [21–23]. Another systematic review came to a rather pessimistic conclusion after focusing on postrandomization exclusions, baseline imbalances, and a more critical analysis of the reporting of patient randomization [24].

These VJs are similar to the determination of alpha or beta errors in designing studies [10,25]. They can be specified a priori, but the exact levels chosen are determined by social convention or personal preference. We must acknowledge that every choice in this regard requires balancing the uncertainty of being wrong in our inferences about study quality with the probability of missing important signals about true benefits and harms from studies of suboptimal quality. The answer to how much uncertainty in study quality we are willing to accept ought to be dependent on the context (e.g., severity of disease, existence of alternatives) and on the preferences and values of the particular patient population to which the evidence will be applied. These VJs, therefore, have to be informed by knowledge in
statistics and clinical epidemiology as well as by knowledge in medicine and ethics. Because there is no “one size fits all” approach for determining how much uncertainty should be tolerated in designing clinical studies or in synthesizing evidence, it becomes important for users of the evidence to be given more information about the investigators' tolerance of uncertainty and their rationale for their choices in a given circumstance.

3. Toward greater transparency

In light of the VJs that shape clinical research and evidence synthesis, a major challenge is to provide greater transparency about these VJs. VJs reflect personal or social preferences about which reasonable people can disagree. Because health care decisions are increasingly subject to public scrutiny and collective influences, it is essential that VJs known to influence production, analysis, and synthesis of evidence be identified and disclosed transparently. The principle of transparency should govern the process of evidence production and synthesis because of the public investment in health care and use of evidence to guide decisions in that investment.

Although it might seem impossible to achieve full transparency about all potential VJs in evidence production and synthesis, a degree of greater transparency could be achieved quite easily. For instance, authors of systematic reviews and health technology assessments, the compilers and synthesizers of evidence, should not only state the most important VJs influencing their reviews but also highlight how their VJs differ from the stated or implicit VJs of previously published systematic reviews on the same topic. Disclosing and justifying choices related to the three key VJs outlined here would take modest effort and should improve the social value of evidence produced for health care decision making. For example, technical tools to help decision-makers (including clinicians, policy-makers, researchers, and patients, depending on the context) select among discordant systematic reviews have been published [26]. Explicit knowledge about VJs in the context of discordant systematic reviews might further help to make choices among alternative health care interventions. Over time, disclosing VJs may come to explain some of the otherwise frustrating variation in guidelines, and may empower users of synthesized evidence and guidelines, to ascertain which guidelines make the most sense for their purposes given the stated VJs.

Acknowledgments
This work was completed while Dr. Daniel Strech was a visiting scholar in the Department of Bioethics, National Institutes of Health, MD, USA. We would like to thank Dr. Franklin Miller for his critical review of the manuscript.

References