**Abstract**

**Background:** Vignette studies of medical choice and judgement have gained popularity in the medical literature. Originally developed in mathematical psychology they can be used to evaluate physicians’ behaviour in the setting of diagnostic testing or treatment decisions. We provide an overview of the use, objectives and methodology of these studies in the medical field.

**Methods:** Systematic review. We searched in electronic databases; reference lists of included studies. We included studies that examined medical decisions of physicians, nurses or medical students using cue weightings from answers to structured vignettes. Two reviewers scrutinized abstracts and examined full text copies of potentially eligible studies. The aim of the included studies, the type of clinical decision, the number of participants, some technical aspects, and the type of statistical analysis were extracted in duplicate and discrepancies were resolved by consensus.

**Results:** 30 reports published between 1983 and 2005 fulfilled the inclusion criteria. 22 studies (73%) reported on treatment decisions and 27 (90%) explored the variation of decisions among experts. Nine studies (30%) described differences in decisions between groups of caregivers and ten studies (33%) described the decision behaviour of only one group. Only six studies (20%) compared decision behaviour against an empirical reference of a correct decision. The median number of considered attributes was 6.5 (IQR 4–9), the median number of vignettes was 27 (IQR 16–40). In 17 studies, decision makers had to rate the relative importance of a given vignette; in six studies they had to assign a probability to each vignette. Only ten studies (33%) applied a statistical procedure to account for correlated data.

**Conclusion:** Various studies of medical choice and judgement have been performed to depict weightings of the value of clinical information from answers to structured vignettes of care givers. We found that the design and analysis methods used in current applications vary considerably and could be improved in a large number of cases.
Background
Preferences and perceived similarities or differences between choice alternatives can be evaluated using structured vignettes. There are two prominent methods of constructing such models of medical judgements, each with their own literature and set of advocates. These are conjoint analysis, developed in the 1970s to study preference and choice [1], and judgement analysis, also called social judgement theory, developed in the 1950s from Brunswik's lens model [2,3]. The two have developed along very different theoretical lines and have developed somewhat different methodology, although there is considerable overlap. Today there is a large number of marketing applications, where the joint effects of multiple product attributes on product choice have been studied. The types of choices include 'ranking', 'rating', and 'discrete choice'.

These methods can be carried forward to the analysis of medical decision making, as medical decisions require judgement under uncertainty. This uncertainty may concern a state, such as the presence of illness, the likelihood of future events, such as those in the natural course of an illness, or the likelihood with which such events may be averted, that is, treatment effects. For many years decision-making research has explored physicians' estimation of probabilities given clinical scenarios [4]. However, there have been concerns whether physicians' probability setting leads to consistent ratings [5]. Moreover, cognitive psychological research shows that physicians do not apply probabilities as suggested by decision-making theory but use their own heuristics to decide [6-8].

Studies of medical choice and judgement offer a way to elicit the public's, patients' and caregivers' views on healthcare that circumvents probability statements [9-11]. The technique is gaining widespread use in healthcare and has been applied in different areas for example to establish patients' preferences in the doctor-patient relationship [12], or to determine optimal treatments for patients [13]. Increasingly, discrete choice analyses are being employed to study how physicians weigh clinical information in the diagnostic work-up. In particular, respondents are asked to rank, rate, or choose between simulated clinical cases varying in values of different symptoms along the possibility that this case will have a certain illness or will need a certain treatment. Comparison with the results of clinical studies allows an analysis of potential discrepancies (e.g. undervaluation of signs and symptoms, overvaluation of test results). Moreover, such comparisons with reference data from clinical studies allow linking physicians' behaviour to illness probabilities and therefore allow examining (implicit) decision thresholds.

A considerable number of studies have been published recently. We provide an overview of existing reports, present an inventory of their objectives and methods, and evaluate them using systematic review methodology.

Methods
We defined a study of medical choice and judgement as an investigation in which preferences were elicited in physicians, nurse practitioners or medical students and that allowed the estimation of the relative importance of different characteristics.

Search strategy
We performed electronic searches in Medline, PsychINFO, CINAHL (Ovid*-version), Web of Science (ISI web of Science*) was used to locate studies that cited four key papers [14-17]. The last update search was performed on 25/3/2005. The exact search strategy may be obtained from the authors.

Inclusion criteria
Eligible articles for this review had to infer cue or attribute weighting from answers to structured vignettes and had to report on caregivers' decision making.

Data extraction strategy
We developed a data extraction form based on the assessment of three articles [17-19]. The form contained twelve items describing a study's salient features of context, design and analysis (for details see Table 1).

Besides some study descriptors such as first author and year of publication, we extracted information on the studies' objectives, the clinical problem, who the decision-maker was, the type of decision/preference (diagnosis, treatment, risk, prognosis, diagnosis & treatment, and other), the number of participants and the authors' aims. The objectives were extracted into five categories: description of preferences in one group of caregivers (1), comparison of two or more groups such as different professions or different levels of competence. (2), assessment of the consistency within caregivers with their actual decisions or their direct rating of the attributes (3), assessment of changes in preferences over time, e.g. after attending a course (4), and comparison of caregivers with guidelines (5a), actual patients' preferences (5b), or the findings of one or more clinical studies (5c). We also registered the number of vignettes, the number of attributes of each vignette and the rationale behind the selection of the attributes. Finally, we documented how participants were asked to respond to the vignettes: rating (yes/no, otherwise), ranking, probability estimates, or discrete choice and the way, if any, in which authors accounted for correlated data in the analysis. We extracted this item because observations resulting from these experiments are typi-
<table>
<thead>
<tr>
<th>Author</th>
<th>Year of publication</th>
<th>Clinical problem</th>
<th>Type of decision</th>
<th>Aim + reference</th>
<th>Number of participants</th>
<th>Type of participants</th>
<th>Number of vignettes</th>
<th>Number of attributes</th>
<th>Source of attributes</th>
<th>Type of outcome</th>
<th>Type of analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kirwan</td>
<td>1983</td>
<td>Rheumatoid arthritis</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>17</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Wigton</td>
<td>1986</td>
<td>Pulmonary embolism</td>
<td>1</td>
<td>5 → B</td>
<td>55</td>
<td>4, 1</td>
<td>27</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Smith</td>
<td>1987</td>
<td>Tube feeding</td>
<td>2</td>
<td>2</td>
<td>232</td>
<td>4, 1</td>
<td>12</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Holmes</td>
<td>1989</td>
<td>Hypertension</td>
<td>1</td>
<td>3</td>
<td>98</td>
<td>3</td>
<td>16</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Von Preyss-Friedman</td>
<td>1992</td>
<td>Tube feeding</td>
<td>2</td>
<td>2</td>
<td>141</td>
<td>4, 4</td>
<td>16</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Lee</td>
<td>1994</td>
<td>Surgical patients</td>
<td>2</td>
<td>2</td>
<td>34</td>
<td>4, 3, 2</td>
<td>30</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Harries</td>
<td>1996</td>
<td>Diversity of diseases</td>
<td>2</td>
<td>3</td>
<td>32</td>
<td>4</td>
<td>130</td>
<td>13</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>McKinlay</td>
<td>1997</td>
<td>Breast cancer</td>
<td>4</td>
<td>1</td>
<td>128</td>
<td>4</td>
<td>32</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Lee</td>
<td>1997</td>
<td>Bile duct stones</td>
<td>4</td>
<td>2</td>
<td>624</td>
<td>4, 4</td>
<td>27</td>
<td>8</td>
<td>1</td>
<td>1, 2, 4</td>
<td>1</td>
</tr>
<tr>
<td>Skaner</td>
<td>1998</td>
<td>Heart failure</td>
<td>1</td>
<td>1</td>
<td>27</td>
<td>4</td>
<td>40</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Timmermans</td>
<td>1999</td>
<td>Colonic emergency</td>
<td>2</td>
<td>2</td>
<td>102</td>
<td>4</td>
<td>16</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Lee</td>
<td>2000</td>
<td>Unstable angina</td>
<td>2</td>
<td>2</td>
<td>18</td>
<td>4, 4</td>
<td>12</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ross</td>
<td>1999</td>
<td>Depression</td>
<td>4</td>
<td>1</td>
<td>407</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Backlund</td>
<td>2000</td>
<td>Hypercholesterolemia</td>
<td>2</td>
<td>5 → A</td>
<td>38</td>
<td>4</td>
<td>40</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Haggerty</td>
<td>2000</td>
<td>Fetal risk situation</td>
<td>3</td>
<td>1</td>
<td>573</td>
<td>2</td>
<td>32</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Skaner</td>
<td>2000</td>
<td>Heart failure</td>
<td>1</td>
<td>5 → C</td>
<td>70</td>
<td>4, 4, 1</td>
<td>40</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Bouma</td>
<td>2001</td>
<td>Aortic stenosis</td>
<td>2</td>
<td>2</td>
<td>275</td>
<td>4, 4</td>
<td>32</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Engelsbel</td>
<td>2001</td>
<td>Ectopic pregnancy</td>
<td>1</td>
<td>4</td>
<td>27</td>
<td>4</td>
<td>16</td>
<td>6</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Kee</td>
<td>2002</td>
<td>Renal disease</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>50</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Sorum</td>
<td>2002</td>
<td>Acute otitis media</td>
<td>4</td>
<td>2</td>
<td>75</td>
<td>4, 4</td>
<td>46</td>
<td>15</td>
<td>3</td>
<td>1, 2</td>
<td>1</td>
</tr>
<tr>
<td>Sorum</td>
<td>2002</td>
<td>Acute otitis media</td>
<td>4</td>
<td>4</td>
<td>75</td>
<td>4, 4</td>
<td>46</td>
<td>15</td>
<td>1</td>
<td>1, 4</td>
<td>2</td>
</tr>
<tr>
<td>Wahlström</td>
<td>2002</td>
<td>Asthma</td>
<td>2</td>
<td>5 → A</td>
<td>314</td>
<td>4, 4, 4</td>
<td>18</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Bouma</td>
<td>2004</td>
<td>Aortic stenosis</td>
<td>2</td>
<td>5 → B</td>
<td>34</td>
<td>4</td>
<td>32</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Sorum</td>
<td>2003</td>
<td>Prostate cancer</td>
<td>4</td>
<td>2</td>
<td>65</td>
<td>4, 4, 4</td>
<td>32</td>
<td>5</td>
<td>1</td>
<td>1, 2</td>
<td>2</td>
</tr>
<tr>
<td>Tamayo-Sarver</td>
<td>2003</td>
<td>Opioid analgesic</td>
<td>2</td>
<td>1</td>
<td>2872</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Mays</td>
<td>2004</td>
<td>Vaccine program</td>
<td>2</td>
<td>1</td>
<td>224</td>
<td>2</td>
<td>13</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Raley</td>
<td>2004</td>
<td>Papilloma vaccine</td>
<td>2</td>
<td>1</td>
<td>181</td>
<td>4</td>
<td>13</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Arnold</td>
<td>2005</td>
<td>Resp. tract infection</td>
<td>2</td>
<td>1</td>
<td>257</td>
<td>4, 4</td>
<td>16</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Lee</td>
<td>2005</td>
<td>Postoperat. recovery</td>
<td>5</td>
<td>1</td>
<td>60</td>
<td>4, 3, 2</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Tielemier</td>
<td>2002</td>
<td>Depression</td>
<td>2</td>
<td>5 → A</td>
<td>449</td>
<td>4, 4, 4</td>
<td>22</td>
<td>7</td>
<td>3</td>
<td>4, 6</td>
<td>1</td>
</tr>
</tbody>
</table>

1) 1 = diagnosis, 2 = treatment, 3 = risk, prognosis, 4 = diagnosis & treatment, 5 = other
2) 1st digit describes aim: 1 = descriptive, 2 = group comparison, 3 = consistency, 4 = change over time, 5 = comparison with reference
3) 2nd digit describes reference: A = guidelines, B = actual patients, C = clinical study
4) 1 = literature, 2 = patients, 3 = expert, 4 = guidelines, 5 = no information
5) 1 = probability, 2 = rating, 3 = ranking, 4 = yes/no choice, 5 = discrete choice, 6 = >2 alternatives
6) 1 = no adjustment for correlated data, 2 = adjustment for correlated data
cally not independent. Each respondent evaluates each of the vignettes. This makes the data from one respondent more alike than one would expect under the assumption of independence, and therefore standard deviations of the attributes could be underestimated. We searched for any statistical method that allows to adjust the standard errors for the intra-group correlation.

All studies were assessed in duplicate. Discordant scores based on reading errors were corrected. Discordant scores based on real differences in interpretation were discussed and resolved through consensus.

**Results**
The searches retrieved 2001 records. Full papers of 81 potentially relevant studies were obtained. In total 51 articles did not meet the inclusion criteria and were excluded after reading the full reports, leaving 30 reports published between 1983 and 2005 for evaluation. (See flowchart in the Figure 1) The salient features of included studies are shown in the Table 1.

**General aspects**
Although the first study was published in 1983, 24 studies (84%) were published after 1995.

Twenty-seven out of thirty studies examined decision behaviour of medical experts [15,17-42]. In half of the studies more than one type of respondent was surveyed. Twenty-eight different medical problems were addressed. Twenty-two (73%) studies examined treatment decisions. Eleven studies (37%) asked for a preferred diagnostic decision, sometimes (6 studies) in combination with a treatment decision.

**Objectives**
Ten studies (33 percent) aimed at describing decision preferences of specific groups of participants [20,24,26,27,29,30,32,38,43,44] and nine studies (30 percent) described decision preference differences between groups [18,25,28,31,34,36,37,40,41]. Three studies explored the consistency of decisions between groups of experts [15,23,45] and two studies examined change of preferences after an intervention [22,35]. Only six studies (20 percent) compared decision behaviour against some sort of empirical reference such as a guideline [21,39,42] (n = 3), actual patient data [17,19] (n = 2) or the result of a clinical study [33].

**Design**
The median number of attributes was 6.5 (inter quartile range IQR 4–9, range 2–15). In 20 studies (67%) the selection of attributes was based on information like the literature [17,20,27,28,31,32,34,35,37,40,43,45] (12 studies), expert opinion (7 studies) or guidelines (1 study). In five studies patient files [15,21,24,33,41] were used to construct the vignettes.

The median number of vignettes was 25 (IQR 16–32), ranging from 3 to 130.

Authors used several response modes for the vignettes. In eight cases they used more than one response mode. In 23 cases authors used a rating procedure [15,18,20-25,27-31,34-37,40-45], where respondents had to rate the relative importance of a given vignette or assign a probability (n = 6) to a diagnosis or outcome [31-33,35-37]. One study used a ranking design, where respondents had to arrange each of the attributes in descending order of importance [24]. In six studies respondents could reply with a yes/no choice [27,30,31,35,38,39]. One study used a conventional discrete choice mode, where respondents, given two or more vignettes, had to select one with the highest likelihood of postoperative recovery [26].

**Analysis**
Twenty (67%) studies did not correct for correlated data. Consequently, only ten studies applied some statistical procedure to account for this correlation within the data [15,22,26,32-37,41].

**Discussion**
This review has two main findings. First, studies of medical choice and judgement are regularly used in the medical field to explore healthcare providers' decision behaviour or preferences. Second, we found a broad spectrum of different methods, and both design and analysis were suboptimal in some cases.

**Cognitive burden/complexity**
One fourth of our studies either contained vignettes with more than nine attributes or compiled sets of over forty vignettes in the same experiment. Empirical evidence showing that these figures are too high is scarce and there is much controversy particularly about the number of vignettes [46]. From a cognitive psychological point of view both figures appear to be very high and could bias the results. This bias typically occurs because respondents are able to integrate and process large information quantities provided simultaneously, or because respondents lose attention when sifting through too many vignettes. However, evidence suggests that more attributes, more choice options and more vignettes decrease response reliability, but do not bias mean responses [46]. As a rule of thumb, the number of attributes per vignette should not exceed six to eight [47-49]. There is much opinion and controversy about maximally allowed number of vignettes, but little rigorous evidence [46]. A re-analysis of 21 commercial studies
suggests a maximum of 20 vignettes [48] and a review of discrete choice experiments evaluating healthcare shows that the number of vignettes seldom exceeds 16 [49]. Furthermore, the majority of studies either used a ranking or rating response mode. These two modes imply very strong assumptions about human cognitive abilities making it more likely that measures will be biased and invalid [50]. Consequently, we therefore recommend the choice based approach.

**Validity, usefulness of study objectives**

In contrast to applications in marketing research where the main topic of a study is to identify opinions regarding a new product, we would be particularly interested to learn about the correctness of care givers’ weighting of the value of clinical information in decisions. While there is no normative benchmark for a “correct” product there is usually one in medical judgement if clinical studies are available. For example, if the results of a study on medical choice and judgement showed that physicians consistently attribute high weights to relatively uninformative lab test but instead undervalue the informativeness of cues from clinical examination they would hint at something that needed to be improved perhaps with an educational intervention. Also the method would allow assessing the change in preferences after intervening with educational measures.

---

**Figure 1**

Study flow.
Most studies did not compare the attributed weights to some sort of normative benchmark such as the results of a clinical study. We only found one out of 30 studies that actually examined this and another five that used a further normative reference (guidelines or patient files). In absence of a normative benchmark these studies leave it to the reader to approve or disapprove the results. Moreover, assessment of discrepancies between different groups of participants has the problem that these could be explained by different clinical circumstances or other factors rather than group specific differences. On the other hand there are medical situations in which views about optimal choices are controversial. In these situations studies that do not compare caregivers’ decision behaviour (or preferences) to some norm may still be useful in that they allow the examination of present opinions.

**Statistical model**
The majority of studies did not account for correlated data in the analysis. Correlated data occur because each respondent assesses different vignettes. Not accounting for this leads to too small estimates of the standard deviations for an attribute and can mimic a statistically significant association where in fact there is none. Unfortunately, guidelines on the conduct of conjoint analyses have not yet reached consensus about the optimal way to analyse correlated data.

**Limitations**
What are the limitations of this review? We think that the search and appraisal procedures were reliable. However, sometimes classifications were difficult to make because of unclear descriptions in the article. We did not contact authors to clarify these uncertainties. Second, there have been two prominent methods of constructing linear models of medical judgements, each with their own literature and set of advocates. These are conjoint analysis, developed in the 1970s to study preference and choice[1], and judgement analysis, also called social judgement theory, developed in the 1950s from Brunswik’s lens model[2,3]. In this review we did not make a distinction between the two methods because there is substantial overlap in methodology. Arguably this is a weakness of our study. However, since we were interested in providing an overview of all studies that examined medical decisions of care givers using cue weightings from answers to structured vignettes applying all sorts of different methods, we feel that our approach has its own merit.

**Future research**
Our review indicates that current applications of conjoint and judgment analysis in the medical field remain suboptimal in some instances. We think that researchers should consider our propositions to ensure internal validity. Moreover we believe that studies investigating care givers’ judgements are most valuable if they allow comparisons with some norm and if they include an assessment of deviations from that norm. Our review only found few such investigations. From a more methodological point of view we agree with a statement in a recent editorial that research is required to learn whether individuals do behave in reality as they state in a hypothetical context. [51]

**Conclusion**
We believe that studies of medical choice and judgement offer many attractive and new insights into medical action. Provided that both methods and application evolve they offer a unique opportunity to improve quality of care.

**Competing interests**
The authors declare that they have no competing interests.

**Authors’ contributions**
AGHK conceived of the study and LMB obtained funding. AGHK and LMB designed the study, supervised the work and drafted the manuscript. AM and AB carried out the data extraction. AM, AB, UH and GtR participated in the design of the study and gave important conceptual input. All authors read and approved the final manuscript.

**Acknowledgements**
We thank Dr. Pius Estermann (information specialist, University Hospital Zurich) for doing the literature searches. This work was supported by the Swiss National Science Foundation (grants no. 323380-103182 and 320080-103183). The funding body played no role in study design; in the collection, analysis, and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

**References**


Pre-publication history
The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/8/50/prepub