International Journal for Quality in Health Care 1998; Volume 10, Number 3: pp. 247–260

Methodology matters – VIII

'Methodology Matters' is the title of series of intermittently appearing articles on methodology. Suggestions from readers of additional topics they would like to see covered in the series are welcome.

Eliciting expert opinion using the Delphi technique: identifying performance indicators for cardiovascular disease

SHARON-LISE T. NORMAND[1], BARBARA J. McNEIL[1], LAURA E. PETERSON[2] AND R. HEATHER PALMER[2]

[1]Department of Health Care Policy, Harvard Medical School and [2]Center for Quality of Care Research and Education, Department of Health Policy and Management, Harvard School of Public Health, Boston, USA

Abstract

Combining opinion from expert panels is becoming a more common method of selecting criteria to define quality of health care. The Rand Corporation pioneered this method in the 1950s and 1960s in the context of forecasting technological events. Since then, numerous organizations have adopted the methodology to develop local and national policy. In the context of quality of care, opinion is typically elicited from a sample of experts regarding the appropriateness or importance of a medical treatment for several well-defined clinical cohorts. The information from the experts is then combined in order to create a standard or performance measure of care. This article describes how to use the panel process to elicit information from diverse panels of experts. Methods are demonstrated using the data from five distinct panels convened as part of the Harvard Q-SPAN-CD study, a nationally-funded project whose goal is to identify a set of cardiovascular-related performance measures.

The development of performance measures or other standards of care involves combining information from several different sources: information from the scientific literature, information that has been synthesized into guidelines, and information from expert groups. This article focuses on the last of these information sources: how to elicit opinion from experts and, once obtained, how to combine the results to make decisions. In the context of developing quality of care, many researchers have employed group judgment techniques in order to make recommendations regarding appropriate treatment. For example, the Agency for Health Care Policy and Research (AHCPR), the National Cancer Institute, and many other organizations in the USA have published guidelines on consensus care. The discussion by Brook (and references therein) in McCormick, Moore, and Siegel [1] provides a particularly informative and concise summary of utilizing group judgments in order to develop medical guidelines. Several articles have appeared in this Journal that apply the Rand procedures [2,3] including the study in this issue by van Berkestijn *et al.* involving non-acute abdominal complaints [4]. From a data-analytic perspective, many researchers have discussed how to combine expert opinion data, using linear regression models for interval-valued data [5], generalized estimating equations for correlated ordinal response data [6,7], and Bayesian models for ordinal responses [8,9].

The goal of this article is to review briefly the methods of elicitation and the subsequent analysis of group judgments by examining an example in detail. Eight national studies funded by the AHCPR are currently underway, focusing on the development and implementation of measures of quality in large-scale settings. The set of studies are denoted Q-SPAN (Expansion of Quality Measures) project. The example considered in this article utilizes results from one such project awarded to Harvard Medical School, called 'Expansion of Quality Measures for Cardiovascular Diseases', henceforth denoted Q-SPAN-CD. A description of the Q-SPAN-CD project concludes the introduction. A brief review of the Delphi technique, a group judgment elicitation process, is

Address correspondence to Sharon-Lise Normand, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA. Tel: (+1) 617 432 3260. Fax: (+1) 617 432 2563. E-mail: sharon@hcp.med.harvard.edu

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

then presented and demonstrated in an application involving judgments of the usefulness of a set of cardiovascular performance measures. Examples of the elicitation instruments are provided as well as of the supporting materials provided to the Q-SPAN-CD panel participants. A discussion of statistical methods for combining the elicited judgments is presented next, including a description of the potential sources of variation and appropriate analytic models. Methods are again demonstrated using the panels convened as part of the Q-SPAN-CD.

The Harvard Q-SPAN-CD project

The main objective of the Q-SPAN-CD is to define quality measures that are well-respected and nationally-based for a set of three inter-related conditions: acute myocardial infarction (AMI), congestive heart failure (CHF), and hypertension. Disease-specific subgroups of clinical experts were responsible for the development of a set of performance measures using information contained in published guidelines as well as the published literature. Performance measures were proposed on the basis of data from several sources: administrative (enrollment, in- and out-patient data), medical (hospital, rehabilitation, and ambulatory records), and patient data (surveys). The investigators integrated expert opinion regarding the usefulness of the proposed measures from several constituencies using a two-stage modified Delphi technique within each of the constituencies. In order to estimate the feasibility, reliability, and usefulness of the performance measures, they are currently collecting retrospective data from sites, representing between a 6-month and 1-year period of care for a sample of patients. The study sites are represented by four participating Health Plans at six delivery settings: Allina Health System in Minneapolis, PacifiCare Health System in Portland, Prudential Health Care in Baltimore and Houston, and United HealthCare in St. Louis and Atlanta. The methodological aspects of the Q-SPAN-CD study featured in this article relate to methods for eliciting expert opinion from health care consumers and suppliers, and statistical methods for combining the expert opinion.

Eliciting expert opinion: the Delphi technique

When empirical data are incomplete or contradictory, expert opinion can often provide a valued resource to aid in decision-making. In fact, study designs that involve groups of experts to elicit judgments have a long history; the voting public and criminal juries are two examples. In the former, each voter may discuss the candidate with whomever he/she chooses but casts his/her vote in anonymity. In the latter, the group discusses features of the case, votes are cast in anonymity, and the votes are counted with the goal of reaching complete consensus. If consensus is not met, more discussion is undertaken and the process is iterated until consensus is reached. The Delphi technique [10] is a method of eliciting group judgment that shares some of the aspects of the two examples. In particular, it is a survey process that is characterized by three features: anonymity, iterative and controlled feedback, and aggregation of responses. These specific features of the Delphi technique are designed to minimize the impact of dominant panelists, discussions not pertinent to the information being elicited, and group pressure towards consensus. The idea is to elicit opinion anonymously through a survey or questionnaire. Often opinions regarding the appropriateness of a medical procedure may be elicited for several different types of patient [11]. The scale of the data collected in the survey is typically ordinal categorical: the medical treatment is 'very appropriate with the benefits far outweighing the risks', 'the benefits equal the risks', 'very inappropriate with the risks outweighing the benefits'. A summary of the results is communicated to the panelists and the sequence is iterated until a stopping point has been reached. The iterations are a device for eliminating noise from the ratings. A two-stage Delphi technique characterizes the process in which two rounds of responses are elicited from the panelists: an initial round of ratings, followed by feedback to the panel, and then followed by a second round of ratings.

Several variations of the Delphi technique are possible. One common modification involves a face-to-face meeting of the panelists after the initial round of ratings (denoted a two-stage modified Delphi technique). Survey items for which there is substantial panel variability are identified and classified as items rated 'with disagreement'. For example, if half of the panelists rated an item at one end of the scale and the other half of the panelists at the other end of the scale, then the item would be considered rated with disagreement. With the aid of a moderator, a meeting is convened and the panelists discuss those items rated with disagreement. At the time of the meeting, each panelist has a statistical summary of the groups' ratings for each item as well as an indicator of the location of his/her response relative to the distribution of responses from the entire group. Clarifications, such as wording, of the survey items are made at this time. Once the discussion has concluded, the panelists confidentially re-rate the survey items for which there was disagreement.

The Q-SPAN-CD advisory panels

In order to develop a set of performance measures that have face validity from the perspective of a wide array of users, advisory panels were created that represented opinion from four key constituencies. Two panels represented views from the buyer perspective: the consumer advisory panel and the purchaser/regulator group. Members of the consumer advisory panel were recruited by the Consumer Coalition for Quality Health Care, Washington, DC and were selected because either they, or someone close to them, had experience with one of the cardiovascular conditions under study. Members of the purchaser/regulator group were recruited by the health plans and consisted of representatives of local business communities, government health programs, and state quality improvement organizations. The remaining two panels provided views from the provider perspective: the physician

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

Table 1 Overview of the advisory panels

	National perspective	Local perspective
Provider perspective	Physician advisory panel	Providers internal to the plans
Buyer perspective	Consumer advisory panel	HMO purchaser and regulatory panel

advisory panel comprised members external to the plans, and a panel of providers internal to the plans. Six specialty societies (American Academy of Family Physicians, American College of Cardiologists, American College of Physicians, American Heart Association, American Medical Association, American Society of Internal Medicine) nomited two practicing clinicians for the Physician Advisory Panel. Each of the participating health plans identified practicing providers from their plans to comprise the providers internal to the plan. By construction, the panel compositions provided national and local perspectives (Table 1).

The two-stage modified Delphi technique

Figure 1 summarizes the Q-SPAN-CD panel process. Group meetings for each individual advisory panel were convened separately by telephone or in person. Panels were run sequentially (consumer advisory panel, then purchaser/regulator group, followed by providers internal to the plans, and then physician advisory panel) with rating information held confidential between the panels. Two rounds of ratings were elicited from each panelist: one prior to each advisory panel meeting and one subsequent to the panel meeting. Prior to the panel meetings, the initial round of ratings was summarized and those performance measures for which there was substantial disagreement among the panelists were identified; the Appendix describes the Q-SPAN-CD algorithm for classifying measures as rated with disagreement. Only those measures that were identified as rated with disagreement in the initial round were re-rated in the second round.

A unique feature of the Q-SPAN-CD project was the formation of a steering committee. This committee was given the ultimate decision-making power with regard to which performance measures were to be initially tested. However, the members were explicitly charged with integrating the data from all the advisory panels with their own expertise in order to arrive at a decision. The committee consisted of investigators from the Harvard Medical School, the Harvard School of Public Health, and the health plans, as well as health plan medical directors and information system directors. Panel sizes ranged from eight to 18 in round 1 and eight to 15 in round 2.

Rating the proposed measures

Each panelist was given a package containing the literature supporting the scientific basis for the set of proposed measures along with a description of each measure. Table 2 displays an example of the descriptive information corresponding to one measure sent to each panelist. The name of the measure, Beta Blocker Prescribed, the clinical rationale for the measure, as well as the definition of the numerator (number who received beta-blockers) and the denominator (number who were eligible to receive beta-blockers) were provided for each of the approximately 100 performance measures. The panelists then answered a series of questions regarding the usefulness along several dimensions such as applicability, likelihood of impacting consumer satisfaction, etc. for each performance measure. Because of the diversity across the panels, the questionnaires were tailored to each advisory panel. Figures 2 and 3 present the questionnaires that were sent to the consumer advisory panel and to the provider internal panel. For each proposed measure, opinions regarding the usefulness and importance over a set of six selection criteria were elicited from the panelists using a five-point scale. A score of 1 indicated the proposed measure was not useful, a score of 3 indicated moderate usefulness, and a score of 5 indicated that the proposed measure was very useful. Note that the judgments regarding how well consumers understand the measures were elicited from the consumer advisory panel while judgments regarding feasibility and predictive validity of the measures were elicited from the providers. Finally, the panelists were asked to synthesize their beliefs regarding the proposed measures into an overall rating using the same five-point scale.

The steering committee utilized a questionnaire similar to that for the advisory panels, but in addition to the scientific literature, summaries of the distributions of the overall ratings from each advisory panel were also included in their package. Figure 4 displays an example of the type of information that was sent to each steering committee member. The distributions of the ratings from the advisory panels were depicted by boxplots along with the number of panelists who contributed ratings to the measure. The boxplots summarized the median, interquartile range, and outlying values of the overall ratings and thus provided the steering committee with a concise yet informative summary. Referring to Figure 4, the buyers (consumer advisory panel and purchaser/regulator group) had a higher overall rating than did the provider (physician advisory panel and providers internal to the plans) panels regarding the usefulness of beta-blockers. In fact, this trend persisted, with the buyers tending to rate the performance measures higher than the providers.

How to combine expert opinion

Once the group judgments have been elicited, the information needs to be combined in order to answer the study question(s) of interest. In the Q-SPAN-CD study, the goal involved using the steering committee's overall ratings to determine which of the proposed performance measures should be field tested. Note that the expert panel data are composed of repeated ratings from each expert; each panelist rated a set of measures

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

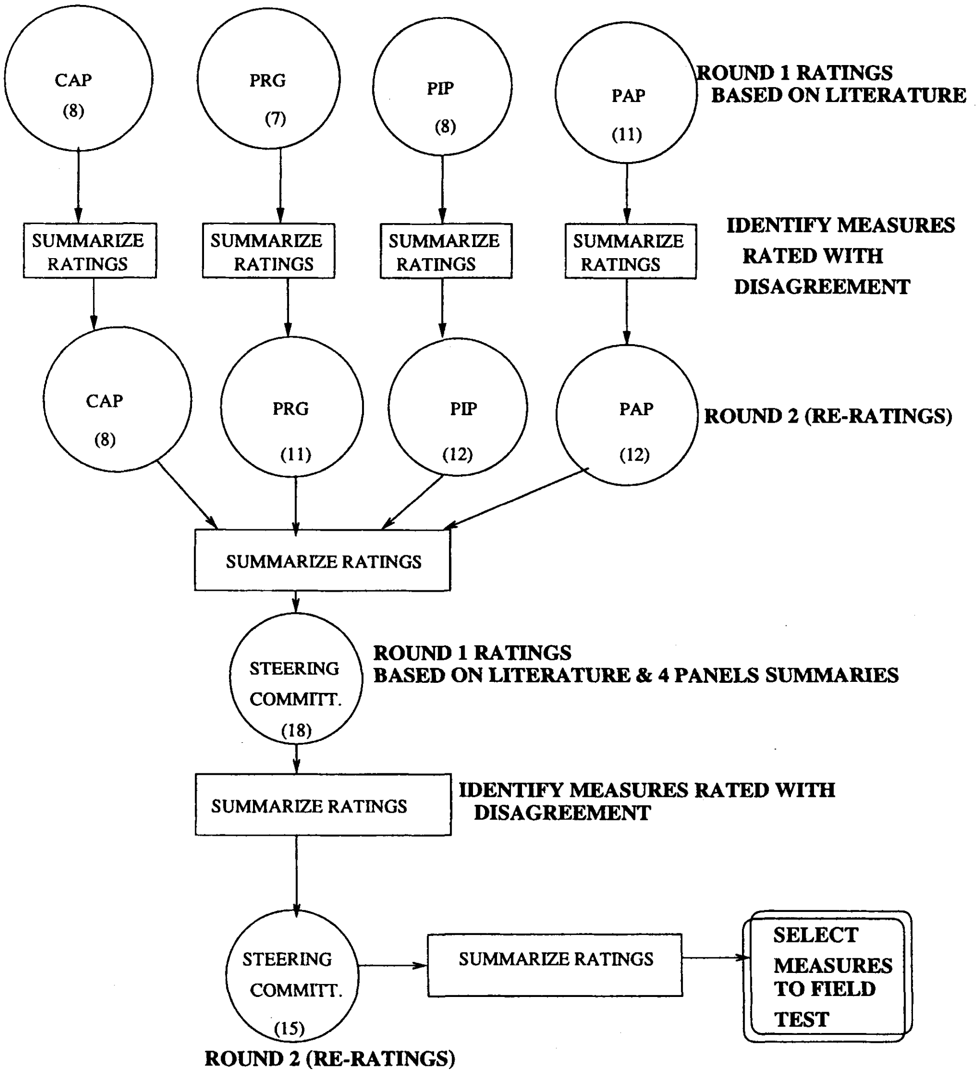


Figure 1 The panel rating process: the four advisory panels independently rated the set of proposed measures using a two-stage modified Delphi process. CAP, consumer advisory panel; PAP, physician advisory panel; PIP, providers internal to the plans; PRG, purchaser and regulator group. In round 1, panelists were sent literature supporting the scientific basis of the proposed measures. Measures for which there was within-panel disagreement were discussed and re-rated in a second round of rating. The round 2 advisory panels' results were summarized and distributed to the steering committee along with the supporting literature. The steering committee rated the performance measures; measures for which there was disagreement among the steering committee panelists were discussed and re-rated in round 2. The committee's final set of ratings was used as a basis to select measures to be field tested. The numbers of panelists are shown in parentheses

using an ordinal categorical scale ranging from 1 to 5. Analytic methods that capitalize on these features are discussed next.

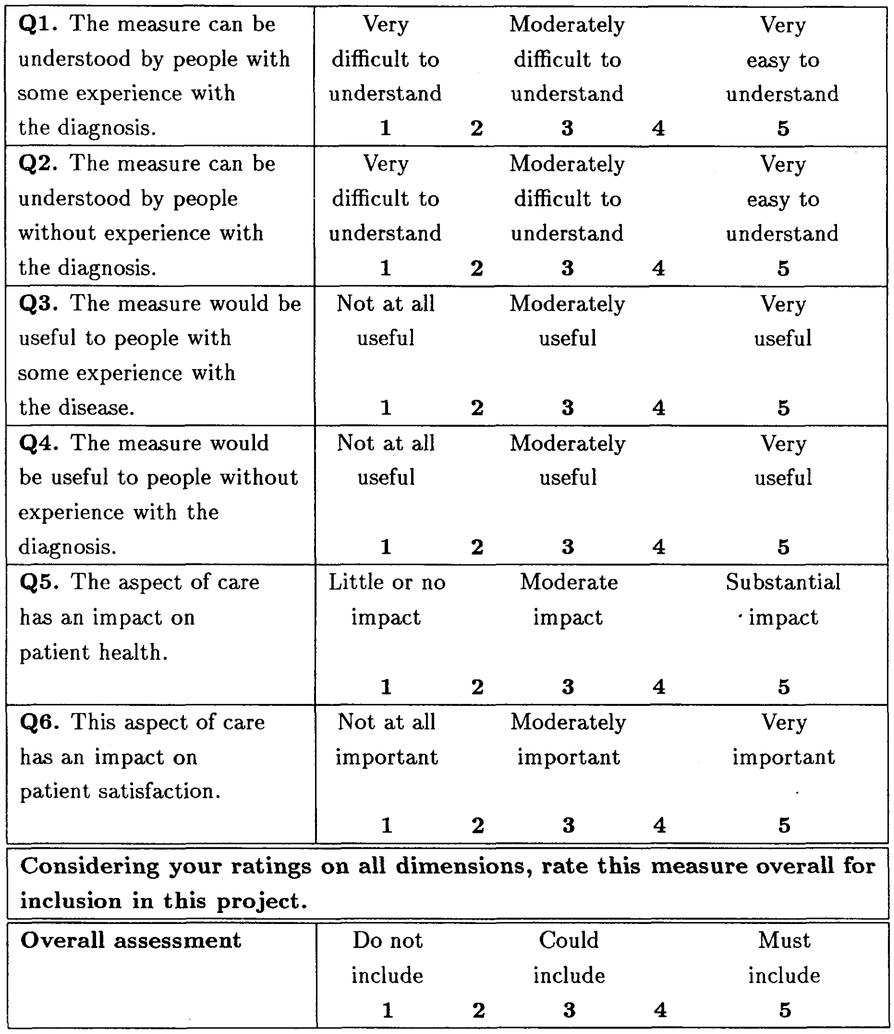
Sources of variability

There are several sources of variability that need to be accounted for when making inferences on the basis of elicited expert opinion. Underlying the panel process is an assumption that there exists a true (unobserved) measure corresponding to each survey item. This unobservable measure is the central quantity of interest and the reason for convening the panels. In many areas, the unobserved quantity is referred to as a latent variable. Using this premise, consider the process that generates a panelist's score for a single performance measurement. There are at least three types of errors to consider: first, there may be a panelist-specific effect. Some panelists may tend to always rate a performance measurement high (or low) relative to the latent variable. Characteristics of the panelist may explain some of the panelist effect; for example, Kahn *et al.* [12] found that beliefs about the appropriateness of coronary angiography differed substantially between primary care physicians, non-invasive cardiologists, and invasive cardiologists. Second, the latent variables may vary across the set of proposed measures. Some

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

Table 2 Description of a proposed performance measure: for beta-blockers in a patient diagnosed with AMI

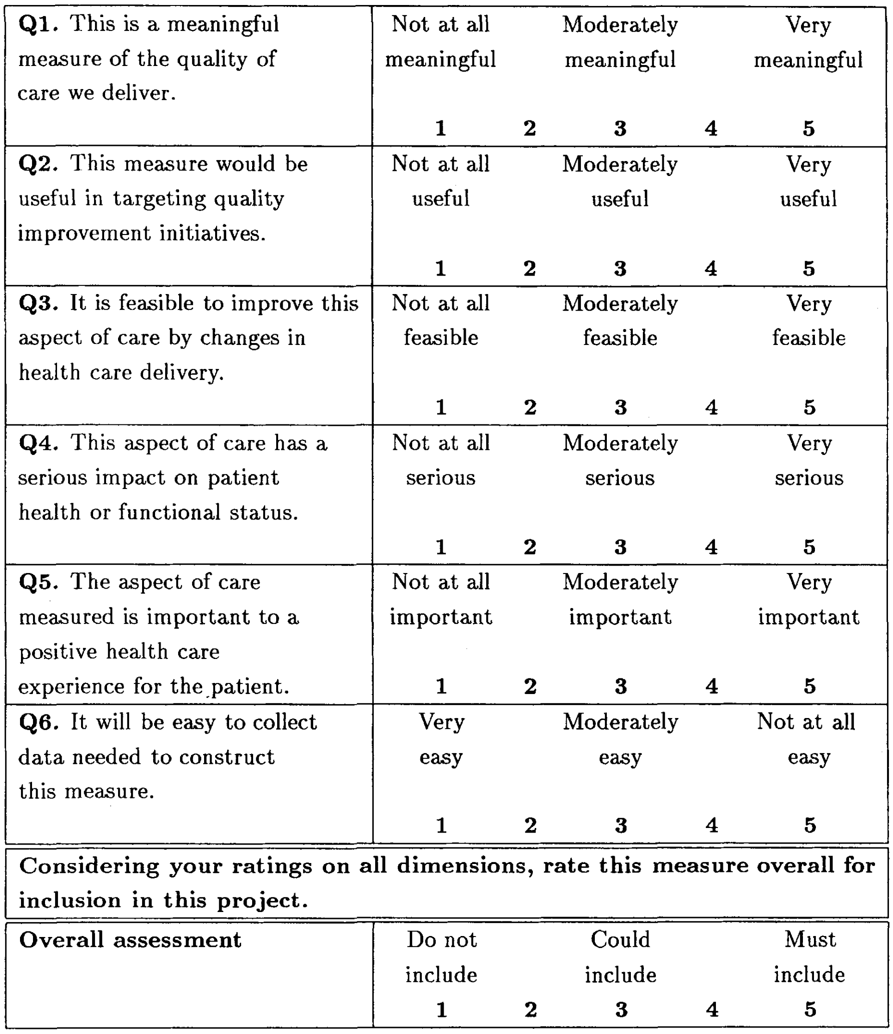
Name of measure	Beta-blocker prescribed
Measure code	AMI-ADM-2.2
Clinical rationale	The use of oral beta-blockers has been shown to decrease morbidity and mortality when prescribed to patients recovering from myocardial infarction Because information on contraindications would not be found in administrative data, this measure does not assess whether a beta-blocker was withheld because of a known contraindication. In Phase I data on contraindications obtained from the medical record will be used to evaluate the extent to which this measure might underestimate actual performance
Denominator	Patients aged ≥30 years who were discharged alive with a principal diagnosis of acute myocardial infarction
Numerator	The number of patients in the denominator for whom the pharmacy database documents a prescription for a beta-blocker between discharge and the end of the first month post-AMI



Q1. The measure can be understood by people with some experience with the diagnosis.	Very difficult to understand **1**	**2**	Moderately difficult to understand **3**	**4**	Very easy to understand **5**
Q2. The measure can be understood by people without experience with the diagnosis.	Very difficult to understand **1**	**2**	Moderately difficult to understand **3**	**4**	Very easy to understand **5**
Q3. The measure would be useful to people with some experience with the disease.	Not at all useful **1**	**2**	Moderately useful **3**	**4**	Very useful **5**
Q4. The measure would be useful to people without experience with the diagnosis.	Not at all useful **1**	**2**	Moderately useful **3**	**4**	Very useful **5**
Q5. The aspect of care has an impact on patient health.	Little or no impact **1**	**2**	Moderate impact **3**	**4**	Substantial · impact **5**
Q6. This aspect of care has an impact on patient satisfaction.	Not at all important **1**	**2**	Moderately important **3**	**4**	Very important **5**
Considering your ratings on all dimensions, rate this measure overall for inclusion in this project.					
Overall assessment	Do not include **1**	**2**	Could include **3**	**4**	Must include **5**

Figure 2 Rating form for consumer advisory panel

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012



Q1. This is a meaningful measure of the quality of care we deliver.	Not at all meaningful 1	2	Moderately meaningful 3	4	Very meaningful 5
Q2. This measure would be useful in targeting quality improvement initiatives.	Not at all useful 1	2	Moderately useful 3	4	Very useful 5
Q3. It is feasible to improve this aspect of care by changes in health care delivery.	Not at all feasible 1	2	Moderately feasible 3	4	Very feasible 5
Q4. This aspect of care has a serious impact on patient health or functional status.	Not at all serious 1	2	Moderately serious 3	4	Very serious 5
Q5. The aspect of care measured is important to a positive health care experience for the patient.	Not at all important 1	2	Moderately important 3	4	Very important 5
Q6. It will be easy to collect data needed to construct this measure.	Very easy 1	2	Moderately easy 3	4	Not at all easy 5
Considering your ratings on all dimensions, rate this measure overall for inclusion in this project.					
Overall assessment	Do not include 1	2	Could include 3	4	Must include 5

Figure 3 Rating form for physicians internal to the health plan

measures will inherently be useful measures while others will be inherently less useful measures. Finally, random error will occur. For example, if the panelist were to rate the measure again under identical conditions, his/her rating might be slightly different.

The sources of error enumerated above reflect a partitioning of the variation in the observed ratings within a panel. When multiple panels are convened, additional sources of variation may emerge. For example, one panel may, on average, rate an identical set of measures higher than another panel. Such additional sources may be reflected as additive or multiplicative components.

The Q-SPAN-CD steering committee ratings

Figure 5 shows the observed ratings corresponding to the overall usefulness of 100 proposed measures elicited from 12 steering committee members who participated in both rounds of the modified Delphi technique. The x-axis depicts the rating scale and the y-axis indicates the number of proposed performance measures. The histograms support the assumption of a panelist effect in that there is variability in the shapes of the histograms among the raters for the identical set of measures. For example, panelists 6 and 11 rated more than 80 of the proposed measures as very useful (a score of 4 or 5) whereas panelist 16 had a more moderate (more bell-shaped histogram) view of the measures.

In terms of between-performance measure variability, Figure 6 displays the distribution of the observed ratings made by the 12 panelists in Figure 5, stratified by 12 randomly selected performance measures (three hypertension, four AMI, and five CHF). The x-axis depicts the rating scale but the y-axis now indicates the number of panelists who rated the particular measure. Eight of the 12 panelists rated the prescription of beta-blockers for AMI patients, displayed in the lower right corner of the figure, as a 'must include' (rating of 5). The usefulness of smoking cessation advice for AMI patients, displayed in the upper left corner, ranged from 'might include' (rating less than 3) to 'must include'.

Figure 7 presents a matrix plot depicting the pairwise relationships among the four advisory panels and the steering committee. Each point within each scatterplot of observations represents the overall rating for a proposed performance

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

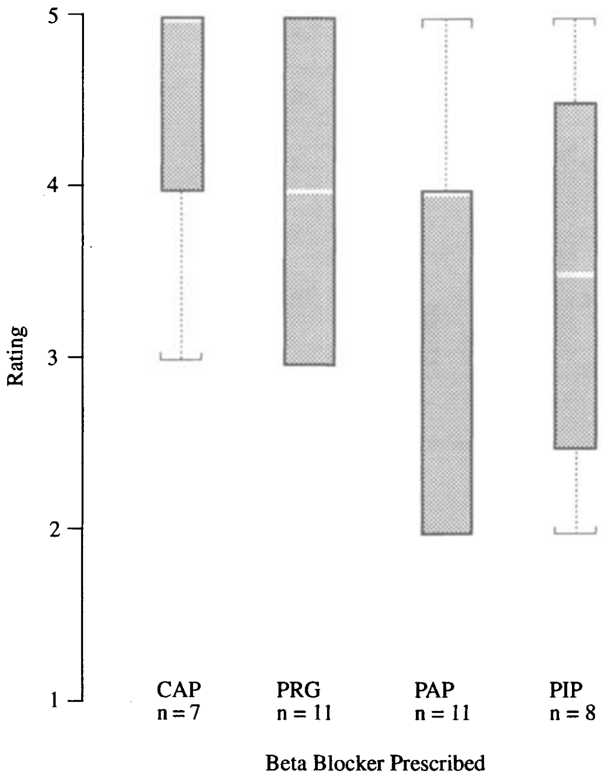


Figure 4 Description of a proposed performance measure sent to steering committee. Steering committee members were sent a book containing descriptions and advisory panel summaries corresponding to 100 performance measures. Below each boxplot, the number of panelists (n) providing ratings is indicated

measure averaged within a panel. Thus, there are 100 points for every pair of panels. A line tracing the average relationship between each pair of panels has been overlaid within each scatterplot to help to show the relationship. First, note that the consumer advisory panel and purchaser/regulator group utilized a smaller range of the rating scale (from 3.5 to 5.0) compared with the average ratings of the other panels. Second, the panels composed of buyers are more similar in their beliefs than to the provider panels. The estimated correlation coefficient between the consumer advisory panel and purchaser/regulator group average rating was 0.28, between the consumer advisory panel and physician advisory panel 0.16, while between the providers internal to plans and physician advisory panel the estimated correlation was 0.76. Similarly, the steering committee ratings are most strongly related to the provider panels' ratings with an estimated correlation between the steering committee and the physician advisory panel overall ratings of 0.88.

Statistical models

There is a tendency to treat expert opinion data as interval-valued data and use an analytic model that assumes the observed ratings are observations from a normal distribution. Unfortunately the scale used to represent group judgment is ordinal categorical, such as 'not at all useful, moderately useful, very useful', rather than interval-valued. In particular, although numerical values are assigned to each category, the distances between categories as measured by the assigned numbers have no real numerical meaning. It is the relative ordering of the numbers that is important; that is, an item rated as a 2 is rated lower than an item rated as a 4. Because the observed ratings are rankings, the distance between a score of 1 and a score of 2 may not be necessarily the same as the distance between a score of 2 and a score of 3. For example, panelists may not differentiate well between 'not at all useful' and 'moderately useful' but have more stringent requirements in moving an item rated as 'moderately useful' to the 'very useful' categorization. Given these features of the expert opinion data, it is good practice to utilize a model that is appropriate for the actual scale of the data [13]. In the case of ordinal data, an analytic model can be built around a regression model by assuming that the panelists implicitly utilize an interval-valued scale, but because the survey response option is constrained to a small number of categories, they choose the category that most closely represents their interval-valued opinion. The ordinal responses are commonly referred to as grouped data since the interval-valued response has been grouped into a small number of categories. In the remainder of this section, models that assume the expert opinion data are interval-valued (ungrouped) and those that assume the data are ordinal (grouped) are discussed.

Analytic models

The observed rating of performance measure i by expert r, denoted $Y_{i,r}$ may be viewed as the sum of three components: the underlying or latent usefulness of performance measure i; an expert-specific deviation (or rater effect) that describes the relative strength of the expert's beliefs about the usefulness of the performance measure; and a measurement error component. Ignoring the scale of the data, a potential model would assume that the observed rating made by expert r for performance measure i is a linear combination of three components:

$$\begin{aligned}&\text{observed rating} = \text{rater effect} + \\ &\text{underlying (latent) score} + \text{measurement error}\end{aligned} \qquad (1)$$

The model described by Equation 1 assumes that the ratings are interval-valued (ungrouped). In order to reflect the fact that the latent variables vary across performance measures but measure a similar construct, each is assumed to be identically distributed and arise from a normal distribution. The rater effects are assumed to be random effects, also arising from a normal distribution, with an average value of zero. Finally, the measurement errors are hypothesized to arise from a normal distribution with constant variance. Models that permit the measurement error structure to vary by rater or by performance measurement may also be estimated.

Alternatively, if the scale of the data is taken into account, then the underlying assumption is that each panelist first

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

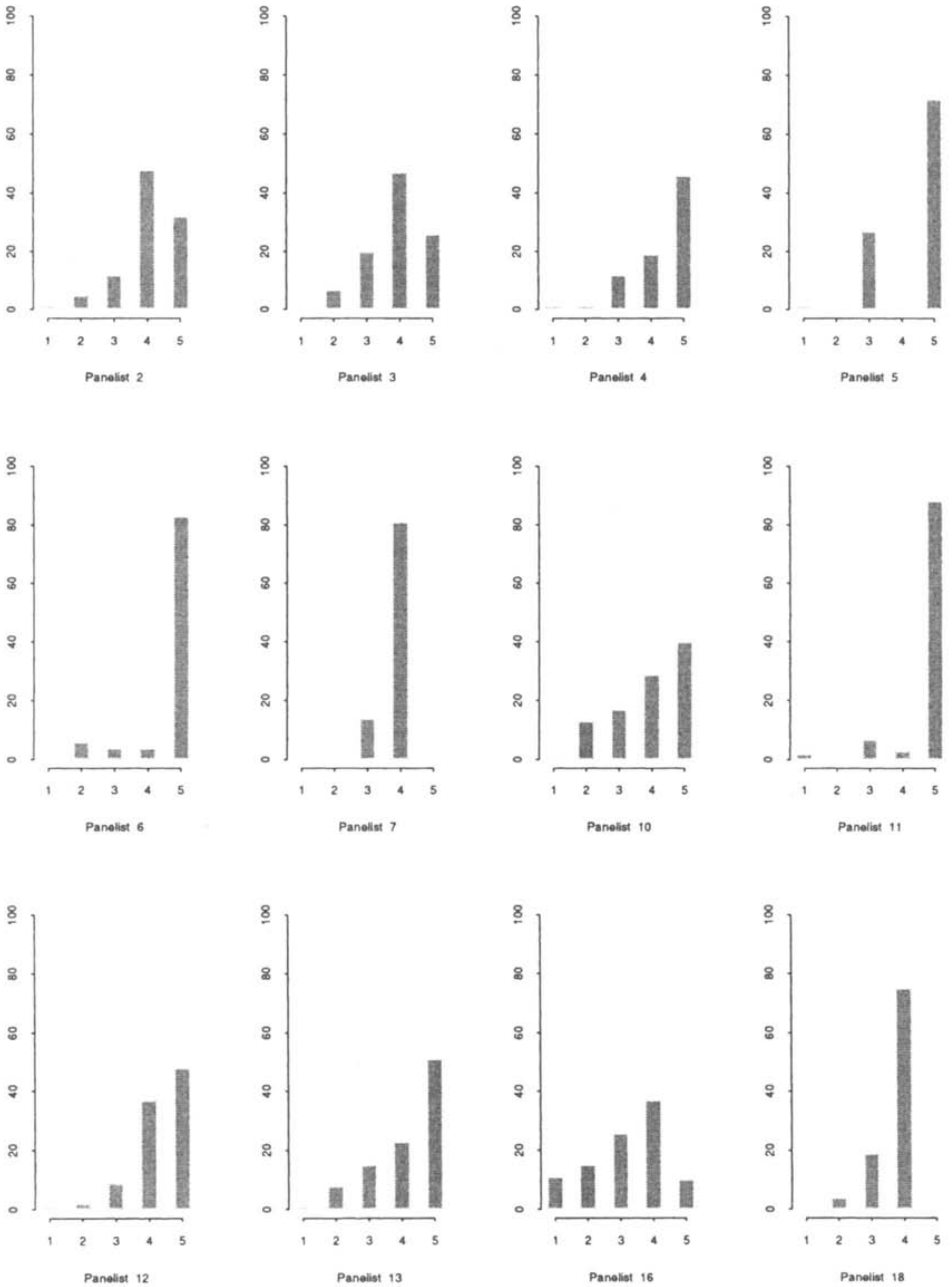


Figure 5 Steering committee round 2 ratings for 100 proposed performance measures. Histograms corresponding to the 12 panelists who participated in both rounds 1 and 2. The x-axis depicts the observed rating for each panelist and the y-axis depicts the frequency of measures. For example, panelist 2 rated approximately 50 of the measures a '4'

generates a rating that is a linear combination of the underlying latent variable and measurement error. Next, the panelist assigns the rating to one of K ordered categories according to his/her own rater-specific threshold values. Thus, for expert r rating measure i, the probability that his/her rating is less than k is such that:

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

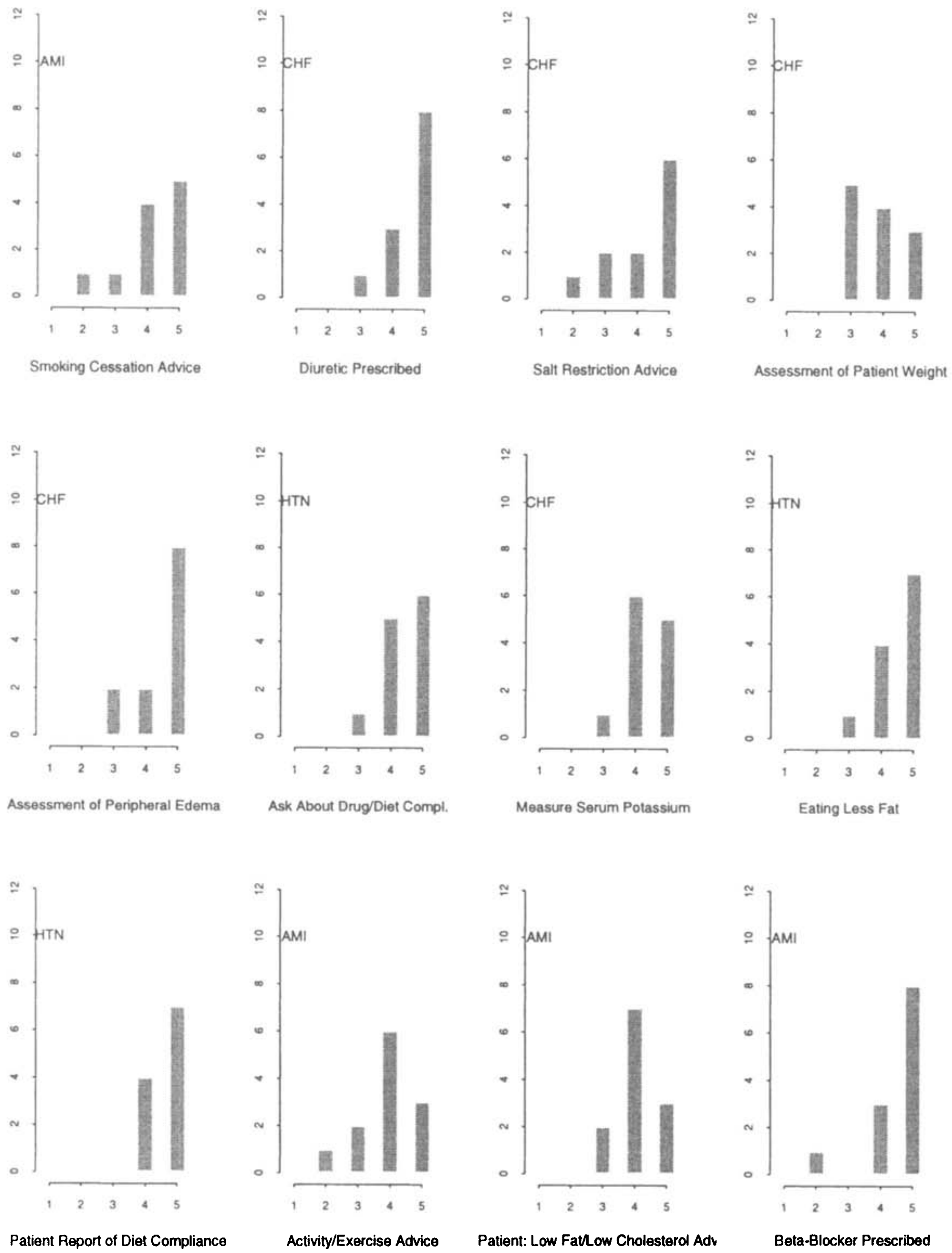


Figure 6 Steering committee round 2 ratings for 12 randomly selected proposed performance measures. Histograms displaying scores for 12 randomly selected performance measures as rated by panelists who participated in both rounds 1 and 2. The x-axis depicts the observed rating for each measure and the y-axis depicts the frequency of panelists. For example, approximately four of the 12 panelists rated 'smoking cessation advice' for AMI patients a '4'

$$P(\text{observed rating} < k) = F\left(\frac{\text{rater-specific threshold for } k - \text{underlying (latent) score}}{\text{standard deviation of measurement error}}\right) \quad (2)$$

where F represents a function. The model described by Equation 2 assumes that the ratings are grouped; that is, ordinal data that are collapsed into K categories and on their original scale are interval-valued. For example, if F denotes

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

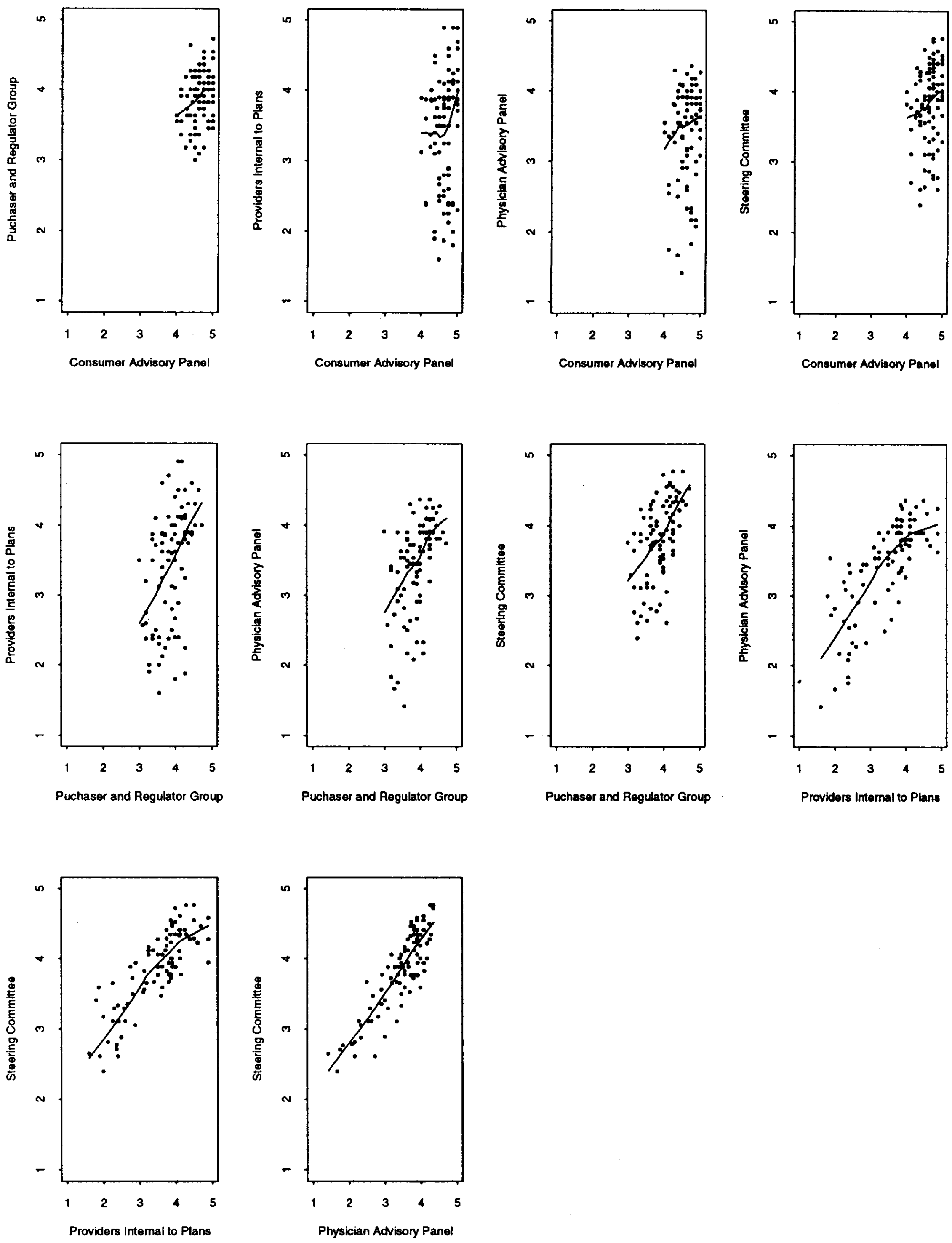


Figure 7 Panel ratings for Q-SPAN-CD performance measures. Pairwise relationship of overall rating, averaged within advisory panel, among panels. The x and y axes display the mean rating per measure, averaged within each panel

the cumulative distribution function for a standard normal distribution, then the model above corresponds to a grouped normal data model. In order to identify the model, the rater effects are scaled to have a standard normal distribution with mean zero as well as subject to order constraints. Similar to the ungrouped models, heteroscedasticity in the measurement error across raters or performance measures may be assumed. Figure 8 shows the latent (continuous) and observed (ordinal categorical) ratings for a fictitious panelist under the assumption that F is the standard normal distribution. The

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

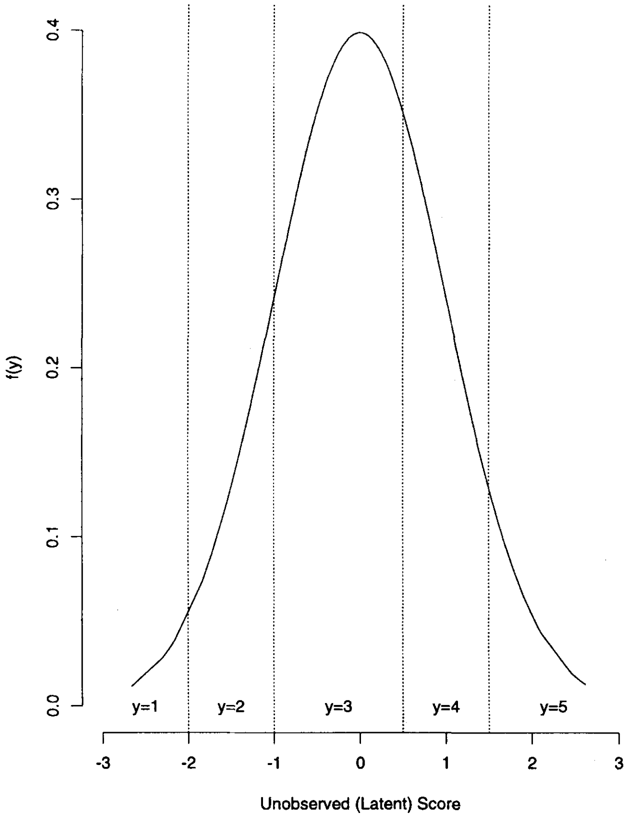


Figure 8 Grouped-normal model for a fictitious panelist. The x-axis displays the unobserved continuous rating (the latent variable), the y-axis displays the probability for each rating, and the dashed vertical lines represent the panelist's thresholds for grouping his/her rating into one of five ordered values. For example, if the latent score is less than −2, the panelist rates the item as y = 1. With five ordered values, four thresholds or panelist-specific effects are necessary

latent ratings are assumed to arise from a normal distribution with mean zero and variance equal to one so that 97.5% of the values of the latent variable lie in the range (−3,3). The dashed vertical lines represent the panelist's threshold values for categorizing his/her latent ratings into one of the five ordered categories. For example, the threshold for classifying an observation into the 'y = 1' categorization is −2. This implies that for this fictitious rater, any (unobserved) score having a value less than −2 would be grouped into the 'y = 1' category. Similarly, if the unobserved score is bigger than −2 but smaller than −1, it would be grouped into the 'y = 2' category by the rater. If we know the threshold values, then the probability that an observed rating is less than k may be easily computed. For example, the probability that $y = 1$ is $\varphi(-2) = 0.0228$ using φ as the normal cumulative distribution function; the probability that $y > 4 = 1 - \varphi(1.5) = 0.0668$.

Modeling the Q-SPAN-CD steering committee ratings

Figure 9 presents a comparison of the estimated rater effects, using either a grouped-normal model or an ungrouped-normal model, for the 18 steering committee panelists who participated in round 2. Each panelist is listed on the vertical axis while the horizontal axis depicts the underlying latent scale. Larger values of the latent score correspond to increasing usefulness of the measure. The dashed vertical lines indicate the quantiles of the estimated distribution of the underlying latent variables. For each panelist, his/her rater-specific threshold parameters, as estimated by a grouped-normal model, are denoted by a solid vertical line. Because there are five ordinal categories, four threshold parameters are estimated for each rater. For example, the mass to the right of the right-most threshold parameter represents the area in which a performance measure would be rated as 'must include' by a panelist. Therefore, panelists with a low threshold value for the right-most parameter would correspond to members who tend to rate measures high relative to the group. The distances between thresholds within a rater are important as well as the between-rater thresholds. Panelist 18, compared with panelist 17, operated on a wider range of the latent score scale in making ratings. Moreover, panelist 18 had a high threshold level for categorizing measures as 'must include'; the threshold value lies beyond the 97.5 percentile of the estimated distribution of latent scores. Note that panelist 5's thresholds for categorizing a measure as a 4 or a 5 were essentially identical. Examination of this panelist 5's observed ratings in Figure 5 indicates, in fact, that no 4's were assigned.

Also included in Figure 9 are the panelist effects corresponding to the ungrouped-normal data model. The estimated effect for each panelist is denoted by a + symbol and often lies between the 2.5 percentile and 97.5 percentile of the latent score distribution. However, in contrast to the grouped-normal model estimated panelist effects (the threshold parameters), the more to the right the estimated effect, the higher (more positive) the panelist effect.

Using the Q-SPAN-CD steering committee ratings to choose measures

The goal driving the panel process involved selecting from the set of proposed measures a subset to be field-tested. In order not to exclude potentially useful measures, the Q-SPAN investigators wanted to eliminate measures that were rated low with *agreement* by the steering committee. Thus, the investigators wanted to minimize their type II error, their risk of rejecting an inherently good performance measure. To operationalize this definition, recall that a rating of 3 corresponded to 'could include' so that the risk that measure i has an underlying rating less than 3 is:

$$P_i = \text{probability underlying score} < 3 \qquad (3)$$

Therefore, if the investigators knew the latent (true) score, they would eliminate measures having scores less than 3. The investigators, *a priori*, subjectively chose 80% as the cutoff, so that if the estimated $P_i > 0.80$, the measure would be

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

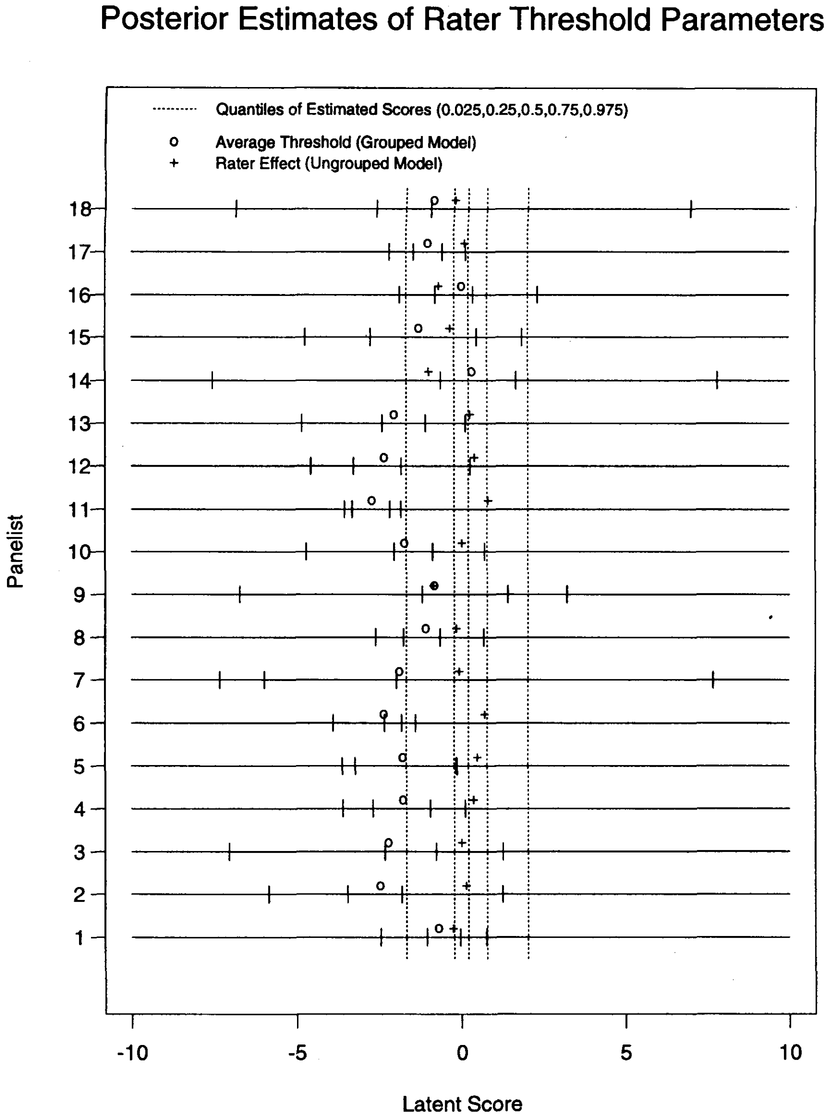


Figure 9 Estimated rater effects: using grouped and ungrouped normal models. The x-axis displays the latent (unobserved) continuous rating scaled to a standard normal distribution. The y-axis displays the rater-specific threshold values for categorizing measures on the one to five scale

eliminated and not carried forward to field testing. Using the model parameters estimated based on the steering committee ratings, $\hat{P}_i$ was constructed for each measure. Figure 10 displays the estimated P_i values on the horizontal axis and the proposed measures on the vertical axis. The estimates were obtained using the group-normal model assuming homogeneous measurement error. No measure had $\hat{P}_i > 0.80$; eight measures have an associated $\hat{P}_i$ greater than 20%. In fact, all measures had $\hat{P}_i$ less than 60%. The largest $\hat{P}_i$ corresponded to a CHF measure involving patient-reported performance of daily weights. Based on the steering committee's ratings and an *a priori* decision rule, no measure was eliminated from the testing set.

Conclusions

Expert opinion can provide valuable information when there is conflicting or incomplete knowledge. In this article, methods for convening the panels as well as statistical models for making inferences on the basis of the panel data were reviewed and demonstrated using results from the Q-SPAN-CD study. One of the unique features of the Q-SPAN-CD is its four diverse constituencies representing local, national, buyer, and provider perspectives. Based on the panel results, there appears to be (preliminary) evidence that the four constituencies can be satisfied by the same set of measures.

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

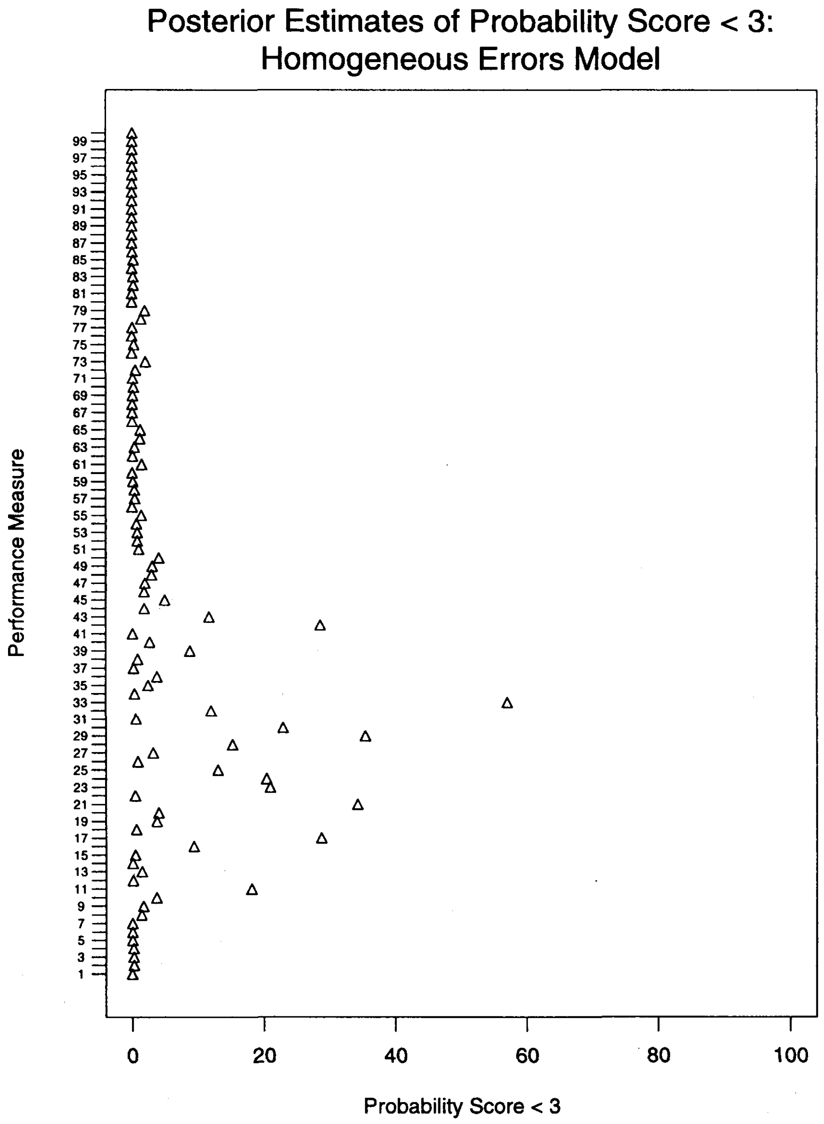


Figure 10 Steering committee overall rating: final round estimated probability (x-axis) that each proposed performance measure (y-axis) is rated low with agreement

Although two classes of analytic models (grouped and ungrouped data models) were presented, models that make proper use of the scale of the data should generally be utilized. In the example considered in this article, the grouped-normal model satisfied this requirement by definition. It is important to note, however, that the computational complexity of the grouped-normal model described in this article is substantially greater than that when estimating the ungrouped-normal model. Assessment of model fit is also challenging and, although it is not discussed here, it is a critical part of any analysis. It is difficult to predict *a priori* how conclusions will differ between the two models without actually estimating both models. The general recommendation is to utilize a model that is commensurate with generation of the data; in this case, models that exploit the scale of the data and the design of the study should be employed.

Acknowledgements

Thanks are extended to additional Q-SPAN-CD Investigators; Department of Health Care Policy, Harvard Medical School: Edward Guadagnoli PhD, Paul Hauptman MD, Daryl Caudry SM, Jeff Souza MA, and Shawna Hutchison BS; Allina Health System: Judy Johnson MHSA, John H. Kleinman MD; Health Care Education and Research Foundation: Catherine Borbas PhD; PacifiCare Health System: Marianne Laouri PhD, Gail Goss RN; Prudential Health Care Research: Nicole Fehrenbach MPP, Jeffrey Koplan MD; and United HealthCare: Eileen Peterson MSPH, Deborah Shatin PhD. We are also grateful to the Clinical Experts, Advisory Panel and Steering Committee participants; and to Mary Beth Landrum PhD, Department of Health Care Policy, Harvard Medical School for computing assistance. This work was supported, in part,

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012

by Grant U18-HS09487 from the Agency for Health Care Policy and Research, Rockville, MD.

References

1. McCormick KA, Moore SR, Siegel RA. *Methodology Perspectives*, AHCPR Pub. No. 95-0009, Public Health Services. Rockville, MD: US Department of Health and Human Services, 1994: 59–70.
2. Bernstein SJ, Hofer TP, Meijler AP, Rigter H. Setting standards for effectiveness: a comparison of expert panels and decision analysis. *Int J Qual Health Care* 1997; **9:** 255–263.
3. Fraser GM, Pilpel D, Kosecoff J, Brook RH. Effect of panel composition on appropriateness ratings. *Int J Qual Health Care* 1994; **6:** 251–255.
4. van Berkestijn LGM, Kastein MR, Lodder A, Melker RA de, Bartelink M-L. How well are patients treated in family practice? Quality consultations for non-acute abdominal complaints. *Int J Qual Health Care* 1998; **10:** 221–234.
5. Fleiss JL. *The Design and Analysis of Clinical Experiments*. Toronto: John Wiley & Sons, 1986.
6. Toledano AY, Gatsonis CA. Ordinal regression methodology for ROC curves derived from correlated data. *Stat Med* 1996; **15:** 1807–1826.
7. Qu Y, Peidmonte MR, Medendorp SV. Latent variable models for clustered ordinal data. *Biometrics* 1995; **51:** 268–275.
8. Johnson VE. On Bayesian analysis of multirater ordinal data: an application to automated essay grading. *J Am Stat Assoc* 1996; **91:** 42–51.
9. Landrum MB, Normand SL. Applying Bayesian ideas to the development of medical guidelines. *Stat Med* 1998 (in press).
10. Dalkey NC. *The Delphi Method: An Experimental Study of Group Opinion*. Santa Monica, CA: The Rand Corporation, 1969.
11. Bernstein SJ, McGlynn, EA, Siu AL *et al.* The appropriateness of hysterectomy: A comparison of care in seven health plans. *J Am Med Assoc* 1993; **269:** 2398–2402.
12. Kahn JP, Park RE, Leape LL *et al.* Variations by specialty in physician ratings of the appropriateness and necessity of indications for procedures. *Med Care* 1996; **34:** 512–523.
13. Agresti A. *Categorical Data Analysis*. Toronto: John Wiley & Sons, 1990.

Accepted for publication 16 March 1998

Appendix: Quantifying disagreement

Within an Advisory Panel, the Q-SPAN-CD investigators considered two types of disagreement measures: an absolute measure that did not depend on the ratings of other panelists and a relative measure that depended on the distribution of the panel ratings. The investigators based their algorithm on the overall rating. Indicators were classified as measured with disagreement by first applying the absolute measure and then the relative measure:

(1) absolute: measures with an observed range of the overall rating of 4 (maximum possible score − minimum possible score) were considered rated with disagreement. This would occur if at least one panelist gave the proposed measure a rating of 1 and if at least one panelist gave the same proposed measure a rating of 5;

(2) remove from the set of proposed measures those indicators rated with disagreement using the absolute rule;

(3) relative: for the remaining measures:

- For each measure *i*, the coefficient of variation (*CV*) across the raters:

$$CV_i = \frac{\text{standard deviation}_i}{\text{mean}_i}$$

  was calculated;

- the observed CV_i values were ordered from smallest to largest;
- measures corresponding to the top 20% of CV_i values were considered rated with disagreement.

Note that this method of classifying measures as rated with disagreement assumes the data arise from a normal distribution.

Downloaded from http://intqhc.oxfordjournals.org/ by guest on May 22, 2012