Assessment of a Novel Hybrid Delphi and Nominal Groups Technique to Evaluate Quality Indicators

Sheryl Davies, Patrick S. Romano, Eric M. Schmidt, Ellen Schultz, Jeffrey J. Geppert, and Kathryn M. McDonald

Objective. To test the implementation of a novel structured panel process in the evaluation of quality indicators.

Data Source. National panel of 64 clinicians rating usefulness of indicator applications in 2008–2009.

Study Design. Hybrid panel combined Delphi Group and Nominal Group (NG) techniques to evaluate 81 indicator applications.

Principal Findings. The Delphi Group and NG rated 56 percent of indicator applications similarly. Group assignment (Delphi versus Nominal) was not significantly associated with mean ratings, but specialty and research interests of panelists, and indicator factors such as denominator level and proposed use were. Rating distributions narrowed significantly in 20.8 percent of applications between review rounds.

Conclusions. The hybrid panel process facilitated information exchange and tightened rating distributions. Future assessments of this method might include a control panel.

Key Words. Quality indicators, Delphi method, Nominal Group technique

The first step in assessing potential quality indicators is to establish face validity (Campbell et al. 2002). An indicator without face validity is unlikely to be utilized, yet the evaluation of face validity often relies on unstructured assessments that suffer from subjectivity, bias, and poor reliability. Consensus-based panel evaluation is one method of establishing the face validity of indicators, and two approaches are widely used: the Delphi technique and the Nominal Group (NG) technique.

The Delphi technique utilizes multiple rounds of independent ratings (e.g., by mailed questionnaire). Although specifics of the method features vary by application (Campbell, Cantrill, and Roberts 2000; Garrouste-Orgeas et al. 2010; Guru et al. 2005), typically a panel of experts independently rates indicators and the ratings are then compiled, summarized, and distributed for

review before another round of ratings. This process is continued until the ratings converge and stabilize. This process allows for a large panel, minimizing the influence of individual panelists, and maximizing reliability. However, because the exchange of opinions and information occurs via written documentation, there is no opportunity for interactive discussion.

The NG technique also utilizes an initial independent rating, followed by the distribution of summarized results. At this point the panel then meets, traditionally in person and in some cases via conference call, to discuss opinions regarding the indicators. Panelists then rerate the indicators independently. This technique is based on the RAND appropriateness method (Jones and Hunter 1995; Murphy et al. 1998b; Campbell et al. 2002; Hutchings and Raine 2006) and has been widely used in quality indicator development (Asch et al. 2001; Asch et al. 2002; Guttman et al. 2006; Mularski et al. 2006; Wang et al. 2006; Smith, Soriano, and Boal 2007; Grunfeld et al. 2008) including the Agency for Healthcare Research and Quality (AHRQ) Quality Indicators (Davies et al. 2001; McDonald et al. 2002, 2008). The NG process allows for efficient information exchange among panelists, which is particularly important when panelists offer unique points of view (e.g., different clinical specialties, types of practice). However, successful facilitation of an in-person or call-based panel limits the size, generally to under 15 individuals. Without effective moderation by the facilitator, one or two individuals can unduly influence the discussion. Because of the small panel size, interpanel reliability is limited (Hutchings et al. 2006; Murphy et al. 1998a).

Both techniques provide advantages over unstructured methods (Murphy et al. 1998a; Campbell et al. 2002). First, structured evaluations attempt to combat cognitive biases in judgment. These biases have been well described by Tversky and Kahneman (1974) and are particularly influential in complex tasks. For example, anchoring bias can occur when panelists set their initial responses relative to the opinion of the group. The Delphi Group and NG methods both require an independent initial rating to anchor opinions based on an individual's own knowledge. Furthermore, structured methods focus the discussion on

Address correspondence to Sheryl Davies, M.A., Center for Primary Care and Outcomes Research, Stanford University CHP/PCOR, 117 Encina Commons, Stanford, CA 94305-6019; e-mail: smdavies@stanford.edu. Patrick S. Romano, M.D., M.P.H., is with the Center for Healthcare Policy and Research, Division of General Medicine, University of California Davis School of Medicine, Davis, CA. Eric M. Schmidt, B.A., Ellen Schultz, M.S., Kathryn M. McDonald, M.M., are with the Center for Primary Care and Outcomes Research, Stanford University, Stanford, CA. Jeffrey J. Geppert, J.D., Med., is with the Centers for Public Health Research and Evaluation, Battelle Memorial Institute, Sacramento, CA.

specific topics pertinent to the underlying validity of the indicator and allow all panelists to have access to similar information before evaluation. Finally, these methods allow for objectively quantifying the results for direct comparisons among indicators to better establish consensual validity.

The relative advantages and disadvantages of these panel techniques for evaluations of indicators have not been studied (although other applications have been examined), and methodological modifications are often not assessed. Some recent studies have assessed the impact of different panel compositions and methods, although most studies examined decisions of “appropriateness” rather than assessments of quality measurement (Hutchings et al. 2006; Hutchings and Raine 2006). We sought to develop a method that could incorporate the advantages of each method, and thereby minimize the disadvantages of each. This paper describes our hybrid Delphi/NG method and presents an evaluation of this method as applied during an assessment of quality indicators.

METHODS

We tested a new hybrid Delphi/NG method in the course of evaluating new applications and definitions of the AHRQ prevention quality indicators (PQIs). The 12 AHRQ PQIs are measures of potentially preventable hospitalizations for chronic and acute conditions, developed originally for area-level assessments. For each indicator, we evaluated the validity of six new applications, specifically, three types of uses (quality improvement, comparative reporting, pay-for-performance) and three denominator levels (area, payer, large provider group). We assumed that quality improvement would only apply to provider groups, and that pay-for-performance would not apply to areas. For three indicators, we also evaluated all three uses for a fourth denominator level, long-term care, during the final round of ratings. In all, the final ratings evaluated 81 combinations of indicators, uses, and denominators (e.g., PQI 1 used for comparative reporting of payers would count as one combination for evaluation by the hybrid panel method).

Panel Formation and Indicator Assessment Process

We started the hybrid panel process with a pool of 174 health care professionals nominated by 24 national clinical organizations. We required all panelists to spend at least 30 percent full-time equivalent with patient care or public health programs in an outpatient setting. We aimed to maximize diversity among panelists, selecting participants from a variety of geographic

Table 1: Panelist Characteristics

<i>Characteristic</i>	<i>Delphi Group (n = 42) (%)</i>	<i>Nominal Group (n = 23) (%)</i>
Gender		
Male	62.8	73.9
Female	37.2	26.1
Urban/rural*		
Urban	32.6	30.4
Suburban	14.0	13.0
Rural	7.0	8.7
Multiple/all areas served	16.3	30.4
Academic affiliation*		
Academic practice	27.9	47.8
Nonacademic practice	34.9	30.4
Underserved population served		
In practice*	46.5	69.6
Funding*		
Public	27.9	34.8
Private and/or nonprofit	20.9	39.1
Multiple sources	7.0	0
Specialty		
Generalist [†]	42.9	34.8
Medical specialty	33.3	43.5
Surgical specialty	14.3	13.0
Nursing and education	9.5	8.7
Research interest [‡]		
Public health	14.6	30.4

*As reported by panelist; percentages do not add to 100% due to lack of reporting by some panelists.

[†]Generalist includes internal medicine, family medicine, geriatric medicine, emergency medicine, and public health physician. Medical specialty includes endocrinology, nephrology, cardiology, pulmonology, and asthma, infectious disease. Surgical specialty includes general surgery, vascular surgery, and urology. Nursing includes nurse practitioner and diabetes outpatient management.

[‡]Research interest assignments were derived from nominating organizations and publicly available information on research interests.

areas, clinical specialties, and practice settings. Eighty-eight nominees met our selection criteria. We assigned 48 to the Delphi Group, 25 to the NG (assigning panelists based on practice characteristics to maximize NG diversity), and 15 to an alternate list. When more than one individual would contribute equally to the diversity of a group, panelist assignment between the two groups was random. We further divided the two groups into four panels: 22 clinicians served in the Delphi Core Panel, 26 in the Delphi Specialty Panel, 10 in the NG Core Panel, and 15 in the NG Specialty Panel (see Table 1).

The first two panels (comprising the Delphi Group) evaluated the indicators by e-mail questionnaires only. The third and fourth panels

comprised the NG, whose members participated in conference calls in addition to completing email questionnaires. Each of the Delphi Group and NG included a core set of panelists with general expertise who evaluated all indicators (core panels: internists, emergency physicians, family physicians, geriatricians, and nurse practitioners) and a group of specialists who evaluated only those indicators pertinent to their particular expertise (specialty panels: e.g., endocrinologists reviewed only diabetes indicators).

The exchange of information between the groups is diagrammed in Figure 1. We provided both groups with a packet of information, including a summary of literature-based evidence, indicator specifications, and area-level rates of hospitalization. First, we asked each group (Delphi and Nominal) to evaluate the indicators, then provided separate summaries of numerical ratings back to each group, with an opportunity to respond with additional comments. Following this initial rating period, we gave NG panelists anonymous commentary and first-round ratings from both groups, then held a total of 6 hours of moderated conference calls with NG panelists. During these calls, the NG discussed opinions about the indicators and potential modifications to the definitions. We summarized the call discussions and presented the summary to both the Delphi Group and NG panels, along with results of empirical analyses, which had been requested by panelists to provide information pertinent for interpretation of the indicators. We then asked panelists from both groups to rerate the indicators. Thus, before the conference call and final ratings, panelists received summarized commentary from both panels in an effort to improve information exchange.

Panelists rated each indicator using a 15-item questionnaire with a nine-point adjectival scale for assessing face validity, bias, potential for gaming, and usefulness for the specified application. We used only usefulness ratings to evaluate the hybrid method in accord with previous work (AHRQ 2002a, 2002b, 2003, 2006).

For each group, we calculated the mean, median, and standard deviation of panelists' usefulness ratings. Following the RAND method (Jones and Hunter 1995; Murphy et al. 1998b; Campbell et al. 2002; Hutchings and Raine 2006), we divided the nine-point scale into three ranges: 1–3; 4–6; and 7–9 to assess the level of agreement. We modified the RAND method to account for the very large panel size. We classified agreement as present if fewer than 15 percent of responses fell outside the three-point range containing the median, and disagreement as present if at least 20 percent of responses fell in each of the two extreme three-point ranges (1–3 and 7–9). This structure matches the percentages in the RAND method. We classified the agreement

Figure 1: Hybrid Delphi and Nominal Group Method

*Specialist panelists rated only indicators relevant to their clinical specialty (e.g., nephrologists only rated diabetes indicators).

level as indeterminate if neither of the above criteria were satisfied. We scored the support level for usefulness as “full support” (median 7–9 without disagreement), “general support with some concern” (median 7–9 with

disagreement), “some concern” (median 4–6.9, regardless of agreement status), or “major concern” (median 1–3.9, regardless of agreement status).

Evaluation of the Hybrid Method

We evaluated the method in three ways: discordance of panel final ratings, change in rating dispersion between initial and final ratings, and a regression of panel assignment on median rating, controlling for indicator and panelist characteristics.

We compared final ratings through Pearson correlation of median ratings and agreement of indicator categorization (κ statistic). We performed mixed model linear regression to assess indicator characteristics (use, denominator level, specific indicator rated) and panelist characteristics (public health interest, panel assignment: core versus specialist, group assignment: Delphi versus Nominal) associated with the mean usefulness rating of the indicators using the hybrid method. Panelists were included as a random effect and all other factors as fixed effects. As the PQIs are traditionally used by public health departments and others measuring area-level hospitalization rates, we created an analytic comparison group of those panelists involved in public health. This group included those nominated by the Medical Care Section of the American Public Health Association and those that self-identified their clinical and/or research interest in public health. Finally, for each usefulness rating, we assessed the change in the distribution of panelist responses between the initial round and final round ratings. In addition to examining percent change, we assessed the statistical significance of changes using Fisher exact test, which tests for changes in central tendency and dispersion. For the purpose of this evaluation, results from each panel were not analyzed separately. We considered all p -values $< .05$ to be statistically significant. We performed all analyses using *SPSS* version 17.0 (Chicago, IL).

RESULTS

Concordance between Groups

Of the 81 combinations of indicators, denominators, and uses assessed in the final round, the Delphi Group and NG provided the same level of support for 45 combinations (56 percent) (Table 2). In no instances did discordance occur at the extremes (i.e., one group expressing a support level of “major concern” while the other expressed “full support”). Of the remaining 36 combinations with discordant levels of support, the NG always provided the more extreme median rating, but in two-thirds of these combinations, the difference in median scores was $\leq$

Table 2: Concordance between Delphi and Nominal Group (NG) on Combinations Rated

	<i>Delphi Full Support</i>	<i>Delphi General Support</i>	<i>Delphi Some Concern</i>	<i>Delphi Major Concern</i>
NG full support	8	2	21 (6)*	0
NG general support	0	0	1 (1)*,†	0
NG some concern	0	0	34	0
NG major concern	0	0	12 (5)*	3

*Numbers in parentheses are the number of instances in that cell where $|\text{Median (Delphi)} - \text{Median (NG)}| > 1$. The median difference between groups was < 1 in all other combinations.

†The support level can only be deemed “General Support with Some Concern” if statistical disagreement exists within the panel.

on the nine-point scale, but spanned the category border. In assessing uses of the indicators, discordance was lower for pay-for-performance than other applications (33 versus 49–53 percent combinations discordant). By indicator, discordance was highest for appendectomy (83 percent discordant), followed by asthma and hypertension (66 percent). The direction of the discordance was consistent within each indicator, such that NG panels rated indicators consistently higher or lower than Delphi panels, across all applications. Overall, the interpanel agreement for support categories was fair ($\kappa = 0.28$, $p < .05$). The Pearson correlation between the panels’ median ratings was 0.71 ($p < .0001$).

Change in Distribution between Initial and Final Ratings

Using Fisher exact test, we identified 57 of the 72 combinations (79 percent) with significant changes in dispersion from the first to last round of ratings. Variance increased for 20 percent and decreased for 80 percent of significant changes. The average reduction in variation was similar between panels (NG: 16 percent, Delphi: 14 percent), indicating that both processes led to similar convergence of panelist ratings. The Delphi panel had less average reduction in variance for quality improvement applications (5 percent) than for comparative reporting (17 percent) or pay-for-performance (21 percent); no such difference was noted for the NG panel.

Characteristics Associated with Panelist Ratings

Results of the linear mixed effects models showed that all indicator characteristics examined were significant predictors of mean usefulness rating, but the group assignment (Delphi versus NG) was not (Table 3). Overall,

Table 3: Characteristics Associated with Usefulness Ratings

<i>Factor</i>	<i>Mean (SD) Usefulness Rating</i>	<i>p-Value*</i>
Panelist characteristics		
Public health interest		
Yes	4.7 (2.2)	NS
No	5.4 (2.1)	
Panel assignment		
Core panel	5.3 (2.1)	NS
Specialty panel	4.8 (2.2)	
Group assignment		
Delphi Group	5.8 (1.8)	NS
Nominal Group	5.6 (2.2)	
Indicator characteristics		
Use		
Quality improvement	5.8 (2.0)	< .0001
Public reporting	5.3 (2.0)	
Pay-for-performance	4.7 (2.2)	
Denominator level		
Area	5.7 (1.9)	< .0001
Payer	4.9 (2.1)	
Large provider group	5.2 (2.1)	< .0001
Indicator (e.g., asthma)		

**F*-test of significance.

generalist practitioners (core panels) rated the indicators more favorably than specialists, and those with expressed interest in public health rated the indicators less favorably than other panelists, but these differences were not statistically significant after clustering by panelist. Panelists rated area-level denominator applications the most favorably. They rated pay-for-performance uses lower than comparative reporting uses of the indicators ($t(712) = 4.8, p < .0001$) and quality improvement higher than both other uses ($t(1815) = 5.7, p < .0001$), regardless of denominator level.

DISCUSSION

We used a new hybrid method to take advantage of the relative merits of the Delphi Group and NG methods. If the method facilitates equal access to information across panels, we would expect to see few differences between groups in the final rating round. In this analysis, we found that the groups rated the indicators with the same recommendation or with a between-panel difference in median scores of less than or equal to one in 85 percent of the

combinations assessed. Furthermore, the correlation in median ratings was relatively strong. Therefore, we conclude that the concordance between the panels was substantial. When between-panel ratings differed, the NG always rated the indicator more extremely (more favorably or less favorably) than the Delphi Group. This could be due to bias from small group size and pull of influential panelists, both known limitations of the NG technique. However, it could also be due to better exchange of information used in this hybrid method, resulting in better discrimination between indicators.

Formal evaluation of interpanel agreement regarding level of support for indicators found only fair agreement. Because panelists rated each proposed application of each indicator on a continuous scale, it is likely that the panelists did not intend the difference between five and six to be different than between six and seven. However, the evaluation of interpanel agreement treats these ratings differently as the latter falls on a category border. Thus, this evaluation focuses on the reliability of the panel method itself, including the post hoc categorization of indicator ratings into support levels. Our finding of fair agreement, as compared with relatively high correlations of mean ratings, suggests that the post hoc categorization method may result in more discrimination than intended by the panelists during the rating process, and that reliability may be compromised when ratings cluster around category transitions.

Several studies have examined the concordance of recommendations across panels, although most examined appropriateness ratings rather than evaluations of quality indicators. Hutchings et al. (2006) compared Delphi panels with NG panels, conducted independently, and found that between-group reliability was higher for Delphi panels than for NG panels. Another study comparing 11 Delphi panels found modest agreement between panels. However, each panel was a single specialty panel rather than the multispecialty panels used in our method. Differences in panelist characteristics in that study may account for the lack of agreement in addition to any inherent reliability limitation in the panel process (Campbell et al. 2004). Five other studies found more substantial agreement between different panel types (68–96 percent agreement), although again the rating consisted of different topics and rating scales than were used in the present study (Tobacman et al. 1999; Brown et al. 2001; Escobar et al. 2003; Washington et al. 2003; Hutchings and Raine 2006; Hutchings et al. 2006; Kadam, Jordan, and Croft 2006).

The goal of our multispecialty approach was to encourage the exchange of information, and our hybrid method allowed for more points of view to be included in the process. Such exchange should theoretically result in the narrowing of rating distributions. We observed significant changes in the

distribution of ratings: about 79 percent of 72 combinations had statistically significant changes in distribution, with the final distribution narrower than the initial distribution for 80 percent of these changes. We also observed that neither panelist characteristics (i.e., public health interest) nor panel assignment (i.e., core panel versus specialty panel) accounted for differences in panelists' usefulness ratings.

Without a comparison group, this study is limited in its ability to assess the relative performance of the hybrid methodology. It is unclear whether results would be similar in traditional panel approaches such as a single Delphi or NG panels. Specifically, it is unclear whether the biggest impact is observed when increasing the number of panelists or increasing information exchange. In addition, this study only tested the hybrid method with one stakeholder group, specifically clinicians. Including other stakeholders such as consumers or policy experts might introduce too much complexity to this method. Future evaluation of the method should involve engaging independent NG and Delphi Group processes to assess the impact of utilizing this hybrid technique.

The assessment of potential quality indicators using consensus methods involves complex cognitive tasks. An individual must assimilate known information as well as new information from the group, judge its relevance to the indicator assessment, and weigh the resulting "data" in making final judgments. Factors such as stereotypes, prejudice and "groupthink," or the tendency for individuals to move their opinions in line with those of the group, can all influence the judgments that "experts" make. Therefore, these factors have the potential to modify the overall assessment of indicators. The task is further complicated in this study by the inclusion of multiple potential uses and denominator levels for each indicator.

With this hybrid method, we attempted to address these complexities in several ways. First, to ensure that each panelist had a similar base of information, we summarized and distributed literature-based evidence reviews before the initial evaluation. In addition, to reduce the cognitive burden required to remember both the evidence base and differing opinions from panel members, we provided all panelists with written summaries of other panelists' comments and conference call discussions. To ensure that panelists weighed all important aspects of measurement, we asked panelists to rate various aspects of indicator performance, including importance, susceptibility to bias, and potential for adverse consequences. These ratings served to prime panelists' overall rating of the indicators. Finally, to address potential issues with groupthink and anchoring bias, all panelists rated the indicators independently before any exchange of group opinions occurred. The final evaluation was also completed individually and anonymously.

While we were unable to evaluate the success of this hybrid method compared with more traditional NG and Delphi methods, the relative burden of running a hybrid panel is minimal compared with the benefits of increased panel size and participation. Although other modifications to the Delphi process have been proposed (Raine, Sanderson, and Black 2005), this new method allows for the inclusion of more stakeholder viewpoints (e.g., additional specialties) and offers the potential of stabilizing indicator ratings by ensuring that each panel benefits from the information and ideas offered by the other panel, with relatively little additional resource expenditures.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: Supported by a contract from the Agency for Healthcare Research and Quality—290-04-0020.

Disclosures: None.

Disclosures: None.

REFERENCES

- Agency for Healthcare Research and Quality (AHRQ). 2002a. *Guide to Inpatient Quality Indicators: Quality of Care in Hospitals – Volume, Mortality, and Utilization*. AHRQ Quality Indicators Technical Report.
- Agency for Healthcare Research and Quality (AHRQ). 2002b. *Measures of Patient Safety Based on Hospital Administrative Data: the Patient Safety Indicators*. AHRQ Quality Indicators Technical Report.
- Agency for Healthcare Research and Quality (AHRQ). 2003. *Guide to Patient Safety Indicators*. AHRQ Quality Indicators Technical Report.
- Agency for Healthcare Research and Quality (AHRQ). 2006. *Measures of Pediatric Health Care Quality Based on Hospital Administrative Data: The Pediatric Quality Indicators*. AHRQ Quality Indicators Technical Report.
- Asch, S. M., E. A. Kerr, P. Lapuerta, A. Law, and E. A. McGlynn. 2001. “A New Approach for Measuring Quality of Care for Women with Hypertension.” *Archives of Internal Medicine* 161 (10): 1329–35.
- Asch, S. M., M. G. Sa’adah, R. Lopez, A. Kokkinis, G. A. Richwald, and D. C. Rhew. 2002. “Comparing Quality of Care for Sexually Transmitted Diseases in Specialized and General Clinics.” *Public Health Reports* 117 (2): 157–63.
- Brown, A. D., M. J. Goldacre, N. Hicks, J. T. Rourke, R. Y. McMurtry, J. D. Brown, and G. M. Anderson. 2001. “Hospitalization for Ambulatory Care-Sensitive Conditions: A Method for Comparative Access and Quality Studies Using Routinely Collected Statistics.” *Canadian Journal of Public Health* 92 (2): 155–9.

- Campbell, S. M., J. Braspenning, A. Hutchinson, and M. Marshall. 2002. "Research Methods Used in Developing and Applying Quality Indicators in Primary Care." *Quality and Safety in Health Care* 11 (4): 358–64.
- Campbell, S. M., J. A. Cantrill, and D. Roberts. 2000. "Prescribing Indicators for UK General Practice: Delphi Consultation Study." *British Medical Journal* 321 (7258): 425–8.
- Campbell, S. M., T. Shield, A. Rogers, and L. Gask. 2004. "How Do Stakeholder Groups Vary in a Delphi Technique about Primary Mental Health Care and What Factors Influence Their Ratings?" *Quality and Safety in Health Care* 13 (6): 428–34.
- Davies, S., J. Geppert, M. McClellan, K. M. McDonald, P. S. Romano, and K. G. Shojania. 2001. *Refinement of the HCUP Quality Indicators*. Technical Review Number 4 (Prepared by UCSF-Stanford Evidence-based Practice Center under Contract No. 290-97-0013) Rockville, MD: Agency for Healthcare Research and Quality.
- Escobar, A., J. M. Quintana, I. Aróstegui, J. Azkarate, J. I. Guenaga, J. C. Arenaza, and I. Garai. 2003. "Development of Explicit Criteria for Total Knee Replacement." *International Journal of Technology Assessment in Health Care* 19 (1): 57–70.
- Garrouste-Orgeas, M., J. F. Timsit, A. Vesin, C. Schwebel, P. Arnodo, J. Y. Lefrant, B. Souweine, A. Tabah, J. Charpentier, O. Gontier, F. Fieux, B. Mourvillier, G. Troche, J. Reignier, M. F. Dumay, E. Azoulay, B. Reignier, J. Carlet, and L. Soufir. 2010. "Selected Medical Errors in the Intensive Care Unit: Results of the IATROREF Study: Parts I and II." *American Journal of Respiratory and Critical Care Medicine* 181 (2): 134–42.
- Grunfeld, E., R. Urquhart, E. Mykhalovskiy, A. Folkes, G. Johnston, F. I. Burge, C. C. Earle, and S. Dent. 2008. "Toward Population-Based Indicators of Quality End-Of-Life Care: Testing Stakeholder Agreement." *Cancer* 112 (10): 2301–8.
- Guru, V., G. M. Anderson, S. E. Fremes, G. T. O'Connor, F. L. Grover, and J. V. Tu. 2005. "The Identification and Development of Canadian Coronary Artery Bypass Graft Surgery Quality Indicators." *Journal of Thoracic and Cardiovascular Surgery* 130 (5): 1257.
- Guttmann, A., A. Razzaq, P. Lindsay, B. Zagorski, and G. M. Anderson. 2006. "Development of Measures of the Quality of Emergency Department Care for Children Using a Structured Panel Process." *Pediatrics* 118 (1): 114–23.
- Hutchings, A., and R. Raine. 2006. "A Systematic Review of Factors Affecting the Judgments Produced by Formal Consensus Development Methods in Health Care." *Journal of Health Services Research and Policy* 11 (3): 172–9.
- Hutchings, A., R. Raine, C. Sanderson, and N. Black. 2006. "A Comparison of Formal Consensus Methods Used for Developing Clinical Guidelines." *Journal of Health Services Research and Policy* 11 (4): 218–24.
- Jones, J., and D. Hunter. 1995. "Consensus Methods for Medical and Health Services Research." *British Medical Journal* 311 (7001): 376–80.
- Kadam, U. T., K. Jordan, and P. R. Croft. 2006. "A Comparison of Two Consensus Methods for Classifying Morbidities in a Single Professional Group Showed the Same Outcomes." *Journal of Clinical Epidemiology* 59 (11): 1169–73.
- McDonald, K., P. Romano, J. Geppert, S. Davies, and B. Duncan. 2002. *Measures of Patient Safety Based on Hospital Administrative Data. The Patient Safety Indicators*.

Technical Review 5. Rockville, MD: Agency for Healthcare Research and Quality.

- McDonald, K. M., S. M. Davies, C. A. Haberland, J. J. Geppert, A. Ku, and P. S. Romano. 2008. "Preliminary Assessment of Pediatric Health Care Quality and Patient Safety in the United States Using Readily Available Administrative Data." *Pediatrics* 122 (2): e416–25.
- Mularski, R. A., S. M. Asch, W. H. Shrank, E. A. Kerr, C. M. Setodji, J. L. Adams, J. Keesey, and E. A. McGlynn. 2006. "The Quality of Obstructive Lung Disease Care for Adults in the United States as Measured by Adherence to Recommended Processes." *Chest* 130 (6): 1844–50.
- Murphy, M. K., N. A. Black, D. L. Lamping, C. M. McKee, C. F. Sanderson, J. Askham, and T. Marteau. 1998a. "Consensus Development Methods, and Their Use in Clinical Guideline Development." *Health Technology Assessment* 2 (3): 1–88, i–iv.
- Murphy, M. K., C. F. Sanderson, N. A. Black, J. Askham, D. L. Lamping, T. Marteau, and C. M. McKee. 1998b. "Consensus Development Methods, and Their Use in Clinical Guideline Development." *Health Technology Assessment* 2 (3): 1–88, i–iv.
- Raine, R., C. Sanderson, and N. Black. 2005. "Developing Clinical Guidelines: A Challenge to Current Methods." *British Medical Journal* 331 (7517): 631–3.
- Smith, K. L., T. A. Soriano, and J. Boal. 2007. "Brief Communication: National Quality-Of-Care Standards in Home-Based Primary Care." *Annals of Internal Medicine* 146 (3): 188–92.
- Tobacman, J. K., I. U. Scott, S. Cyphert, and B. Zimmerman. 1999. "Reproducibility of Measures of Overuse of Cataract Surgery by Three Physician Panels." *Medical Care* 37 (9): 937–45.
- Tversky, A., and D. Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–31.
- Wang, C. J., E. A. McGlynn, R. H. Brook, C. H. Leonard, R. E. Piecuch, S. I. Hsueh, and M. A. Schuster. 2006. "Quality-of-Care Indicators for the Neurodevelopmental Follow-Up of Very Low Birth Weight Children: Results of An Expert Panel Process." *Pediatrics* 117 (6): 2080–92.
- Washington, D. L., S. J. Bernstein, J. P. Kahan, L. L. Leape, C. J. Kamberg, and P. G. Shekelle. 2003. "Reliability of Clinical Guideline Development Using Mail-Only versus in-Person Expert Panels." *Medical Care* 41 (12): 1374–81.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.