Technical Correspondence

Effect of Pooled Comparative Information on Judgments of Quality

Leigh A. Baumgart, Ellen J. Bass, *Senior Member, IEEE*, John D. Voss, and Jason A. Lyman

***Abstract*—Quality assessment is the focus of many healthcare initiatives. Yet, it is not well understood how the type of information used in decision support tools to enable judgments of quality based on data impacts the accuracy, consistency, and reliability of judgments made by physicians. Comparative pooled information could allow physicians to judge the quality of their practice by making comparisons with other practices or other specific populations of patients. In this study, resident physicians were provided with varying types of information derived from pooled patient datasets: quality component measures at the individual and group level, a qualitative interpretation of the quality measures using percentile rank, and an aggregate composite quality score. Thirty-two participants viewed 30 quality profiles consisting of information applicable to the practice of 30 deidentified resident physicians. Those provided with quality component measures and a qualitative interpretation of the quality measures (rankings) judged quality of care more similarly to experts and were more internally consistent compared with participants who were provided with quality component measures alone. Reliability between participants was significantly less for those who were provided with a composite quality score compared with those who were not.**

***Index Terms*—Decision support, judgment analysis (JA), quality assessment, quality improvement.**

I. Introduction

The ability to judge quality is an important skill in many domains including healthcare [1], [2] and education [3]. For example, both graduate medical education and clinical specialty certification agencies mandate that physicians must demonstrate the skill of judging the quality of their clinical practice. They must learn to investigate and evaluate practice data in order to judge the quality of the care they provide [4], [5]. Despite the importance of this skill, many physicians are not trained to make quality judgments and have demonstrated a limited ability to accurately judge their quality of care [6]. Our overall goal is to inform the design of tools that will aid physicians in their ability to investigate and evaluate practice data to judge quality.

Pooled datasets provide one opportunity for supporting quality judgments in healthcare. Electronic medical records (EMRs) and related data repositories can provide the automated collection, processing, and presentation of pooled information, meaning aggregated data derived from a defined population of patients rather than episodic data on a single-patient encounter or experience.

Some effort has been devoted to using absolute quality indicators to gain insight into the quality of care provided to pooled patient populations [7], [8]. However, there remain limitations in the validity and use of these direct indicators due to inconsistent definitions, varying data sources, lack of risk adjustment, lack of evidence to predict better patient outcomes, and timeliness of data [9]–[15].

Using comparative information supports overcoming some of these limitations in quality judgment. For example, in healthcare, comparative pooled information allows physicians to judge their practice by making comparisons with other practices or other specific populations of patients. This comparative assessment requires sequential information acquisition of quality measures or indicators, interpretation of these measures, and aggregation of the measures to form a judgment regarding the quality of care provided.

To support the acquisition of pooled comparative quality information, tools may compute and present quality *component measures*. These measures of quality may include specific structure, process, or outcome measures that relate to care in a certain area for specific populations of patients. For example, the Centers for Medicare and Medicaid Services have created a website that presents quality component measures for different hospitals within a geographical area, such as the process measure of the percentage of patients at each hospital who have been given an influenza vaccine [16]. This allows users to compare the percentages between hospitals to make their own judgment regarding quality of the hospitals.

Another strategy for supporting quality judgments with pooled comparative information is to present quality component measures and to also aid in *interpreting* those quality measures. For example, the Leapfrog Group provides a website with qualitative interpretations of individual quality component measures, such as indicating which hospitals have shown "substantial progress" for a quality measure or which hospitals "fully meet predetermined thresholds" for certain quality measures [17].

For some aspects of care, multiple quality component measures are needed to make a judgment regarding the overall level of quality. To ease the amount of information needed, individual component measures have been *aggregated* into a single composite quality score [18]–[23]. While these composite scores are able to statistically describe the variation between populations of patients, they are sensitive to aggregation rules, and the comprehensibility of them by users has not been fully investigated [24].

Manuscript received September 24, 2014; revised February 3, 2015 and June 8, 2015; accepted July 5, 2015. Date of publication August 5, 2015; date of current version November 12, 2015. This work was supported in part by the National Library of Medicine under Grant T15LM009462. This paper was recommended by Associate Editor M. C. Wright.

L. A. Baumgart is with the Center for Biomedical Research Informatics, NorthShore University HealthSystem, Evanston, IL 60201 USA (e-mail: leigh.baumgart@gmail.com).

E. J. Bass is with the College of Computing & Informatics and College of Nursing & Health Professions, Drexel University, Philadelphia, PA 19104 USA (e-mail: Ellen.J.Bass@Drexel.edu).

J. D. Voss is with the Department of Medicine, University of Virginia, Charlottesville, VA 22904-4132 USA (e-mail: jv4w@virginia.edu).

J. A. Lyman is with the Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22904-4132 USA (e-mail: jl3de@virginia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2015.2459382

Despite the increasing effort to develop and present comparative quality information to aid in judging quality of care, it is not known how the type of information considered affects judgments, particularly where gold standards may be based on experts' judgments of the information. The overall aim of this study was to investigate the impact of different types of pooled comparative information on resident physicians' ability to judge quality of care. Specifically, our objectives were to investigate how the display of quality component measures, interpretations of component measures, and an aggregation of component measures affected 1) *accuracy* of quality judgments measured by comparing the participants' quality judgments to those of experts, 2) *consistency* of judgment measured by the ability of participants to maintain their judgment policy over time, and 3) *reliability* of judgment measured using judgment value range information across multiple participants. Assessment of the overall quality of hypertension care in an ambulatory environment was chosen as the setting for this investigation. Understanding these relationships can help inform the design of quality decision support tools that enable investigation and evaluation of practice data to judge quality.

II. Methods

A. Participants

The study received institutional review board approval. Thirty-two resident physicians from the same academic medical center participated in the study. The resident physicians were all in their second or third postgraduate year and were all familiar with comparative pooled patient data. They all had previously participated in at least six 1-h seminars where they were asked to evaluate the quality of ambulatory care provided to the population of patients under their direct care (i.e., their panel of patients). The number of potential participants consisted of 65, but due to scheduling, only 40–45 are available for participation during any one six-month time period (yielding an effective participation rate of over 70%). They were given a $10 gift card for their time.

B. Lens Model Study Design

We employed judgment analysis (JA) techniques [25] and a lens model [26], [27] variant study design to analyze the participants' quality judgments and to address the three study objectives. To consider both internal (cognitive) and external aspects of judgment, lens model designs have been used in other healthcare-related studies of judgment [28], [29]. Using this type of design, participants are presented with multiple "profiles" and asked to make judgments for each. The information used in profiles in this study are described below and consisted of pooled comparative information related to hypertension care. This design affords statistical modeling at the individual level, providing insight of internal accuracy and consistency for an individual judge. Aspects of the judge's policy including the relationship between the cues and the judgment (i.e., cue utilization) can also be considered. At the group level, agreement between multiple judges can be assessed.

TABLE I
Quality Component Measures to Support Judging Quality of Hypertension Care

Component measure	Description
Goal Blood Pressure (Goal BP)	Percent of patients at or below goal blood pressure (where goal was 140/90 mmHg for patients with a diagnosis of diabetes and 135/85 mmHg for all other patients)
Diabetes Mellitus Medications (DM Meds)	Percent of patients with diabetes (and hypertension) and a positive microalbumum test prescribed either ACE (angiotensin converting enzyme) or ARB (angiotensin receptor blockers) medications
Creatinine Lab Checked (Labs)	Percent of patients with creatinine checked within last 15 months

C. Quality Judgment Task

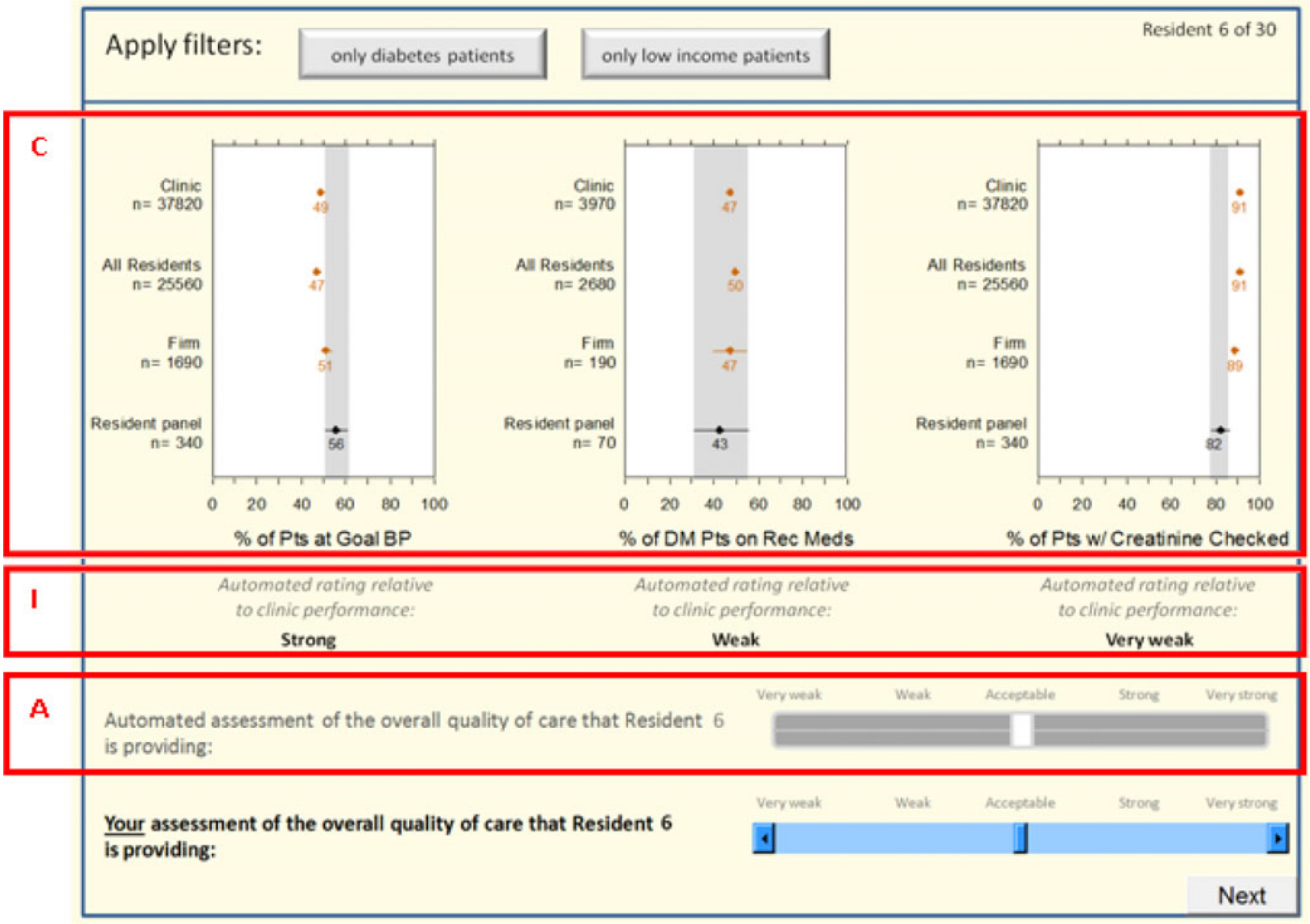
All participants viewed 30 quality profiles consisting of pooled comparative quality information applicable to the practice of 30 resident physicians (all deidentified). Each participant made a quantitative judgment regarding the quality of hypertension care provided by each of the 30 resident physicians under review in each profile.

D. Materials

1) Quality Profile Data: Data applicable to the 30 deidentified resident physicians in each profile were based on both the resident's panel of patients (the population of patients under the resident's direct care) and three additional comparative populations related to the resident: 1) the resident's firm (a term used by the healthcare system to refer to the resident's panel combined with five to six peer resident panels all supervised by the same attending physician); 2) all resident panels in the clinic combined; and 3) the entire clinic combined (including both attending and resident physician panels). The 30 residents used to create the judgment profiles in this study were selected from 103 possible. The criterion for selection was the 30 largest panels with respect to the number of patients meeting the inclusion criteria. These criteria included only patients who had been seen at the clinic within the last year, had a recorded diagnosis of hypertension, and were between the ages of 18 and 74. The average number of patients in the 30 panels used for the profiles was 36 ± 4.

We employed a two-phased method to identify what pooled quality component measures were needed to judge quality of hypertension care. First, a document analysis of clinical guidelines identified five electronically available measures [30]–[34] related to caring for hypertensive patients. Then, a separate focus group of seven internal medicine attending physicians narrowed the measures down to three (based on agreement of a lack of evidence in the literature that two of the measures were indicative of quality hypertension care). Table I indicates the three quality component measures related to hypertension care that were calculated for each population of patients.

Thus, for each of the 30 quality profiles, three types of information were defined to aid participants in judging quality of hypertension care.



Automated rating relative to clinic performance: **Strong**

Automated rating relative to clinic performance: **Weak**

Automated rating relative to clinic performance: **Very weak**

Automated assessment of the overall quality of care that Resident 6 is providing:

Your assessment of the overall quality of care that Resident 6 is providing:

Fig. 1. Quality decision support tool. Types of pooled comparative information (indicated by red boxes overlaying screenshot) include quality component measures (C), interpretation of component measures (I), and an aggregation of component measures (A).

1) The three hypertension quality component measures for the resident under review (i.e., the resident's panel) and for the additional three comparative populations of patients (i.e., the resident's firm, all resident panels combined, and the entire clinic) were available.
2) An interpretation of each quality component measure was also available. This was determined by calculating the percentile rank of the resident's panel compared with all other resident panels.
3) An algorithm was also developed to aggregate the quality component measures into a composite score meant to replicate expert judgment of overall hypertension care. This aggregation algorithm involved two steps. First, a weighted average of the three component measures was calculated. The weighting scheme for this computation was 2:1:1 and was determined by the focus group of internal medicine attending physicians who after agreeing on the three component measures, agreed that the "Goal BP" measure was twice as important as the other two measures. Second, the percentile rank of this weighted average for each resident (compared to all other residents) was calculated, and this percentile was available as the composite quality score.

2) Quality Decision Support Tool: A prototype quality decision support tool to display quality profiles was built in Microsoft PowerPoint and used Visual Basic and ActiveX controls. The tool could display the three different types of pooled comparative information described above: individual quality component measures (C), interpretation of quality component measures (I), and aggregation of quality component measures (A). Fig. 1 depicts the decision support tool displaying all three types (C, I, and A).

For displaying quality *component measures (C)*, the decision support tool presented dot plots, where one plot included one component measure for each of the four populations. In addition, the tool also displayed confidence intervals for all component measures. Confidence intervals were calculated using the Pearson–Klopper method. However, to make the confidence intervals easier to interpret in this study, the populations were scaled by a factor of 10 before calculating the confidence intervals. This scaling did not affect the values of any of the quality component measures as they were always expressed as percentages. Without this scaling, the confidence intervals were large, and pilot testing for this study indicated that some users would disregard the data altogether. Thus, we determined it was an acceptable presentation adjustment for this study if confidence intervals were to remain displayed and not confound the analysis, while noting that future work would need to address this issue.

For displaying *interpretation of component measures (I),* the decision support tool displayed the percentile rank group for the resident under consideration for the quality component measures. If a resident was ranked in the bottom percentile, or

TABLE II
COMBINATIONS OF POOLED COMPARATIVE INFORMATION TO SUPPORT JUDGING QUALITY OF HYPERTENSION CARE

Info Type	Description
C	Dot plots for each quality **component** measure
CI	Dot plots for each quality **component** measure plus the **interpretation** of each component measure using percentile rank groups
CA	Dot plots for each quality **component** measure plus the **aggregation** of the component measures into a composite quality of hypertension care score
CIA	Dot plots for each quality **component** measure, the **interpretation** of each component measure using percentile rank groups, and the **aggregation** of the component measures into a composite quality of hypertension care score

0–20%, for a component measure, "very weak" was displayed, 20–40% was represented with "weak," 40–60% was represented with "acceptable," 60–80% was represented with "strong," and 80–100% was represented with "very strong." A resident under consideration could have different component measure interpretations (e.g., "very weak" for one component measure and "very strong" for another). This scale was chosen to provide more meaning than a simple interpretation of the absolute values of the component measures. For example, if a resident physician only had 60% of their patients at goal blood pressure, but was the highest rank among peers, the interpretation in the form of the rank provides additional context over an interpretation based on the component measure alone.

For displaying the *aggregation of component measures (A)*, the decision support tool displayed the quality of hypertension care composite score as calculated by the algorithm described above. This was indicated with a tick mark along a slider bar.

The quality decision support tool provided two additional functions for users. First, users could filter the data to include *only* patients with a diagnosis of diabetes or *only* patients with low income (defined by eligibility for the hospital's highest level of financial assistance). The filters were added to address concerns indicated in our prior work [35] that comorbidities and socioeconomic factors may bias the interpretation and use of the quality measures. Using the filters allowed for interrogating the measures while holding those factors constant. For both filters, the initial inclusion criteria for patients in all populations were maintained. Second, users could make their quality of care judgments by using a slider bar ranging from 0 (very weak) to 100 (very strong) located on the bottom of the decision support tool.

E. Experimental Design

Participants were randomized into one of four groups (see Table II) based on the type(s) of pooled comparative information available on the decision support tool. The quality component measures (C) were available in all groups. The three combinations of information related to the interpretation (I) and aggregation (A) of component measures were included with the component measures to form the three other experimental groups.

F. Procedure

To begin, all participants were provided with a brief introduction to the study. They then stepped through a self-paced training session that was tailored to their information condition. This training also included three practice quality profiles to allow participants to gain familiarity with the task of judging quality of hypertension care using the prototype decision support tool. Participants were not told how the interpretation (I) or aggregation (A) of the quality component measures was computed.

During the experimental session, participants made 30 judgments of the quality of hypertension care for the 30 profiles (30 deidentified residents). They submitted their judgments using the continuous slider bar (bottom of Fig. 1) displayed for each profile. They were instructed the slider could be moved between the visible qualitative markings. The total time required for the training and experimental session was 30–45 min.

G. Dependent Variables and Data Analysis

Participants' judgments for the 30 profiles were recorded by the decision support tool directly. To address the first objective of investigating the *accuracy* of quality judgments, the residents' judgments were compared with judgments made by five experts. This approach was taken as no gold standard exists for what is considered a "correct judgment" of quality of hypertension care. The experts were internal medicine attending physicians whom had an average of 19.7 years of clinical experience (range 9–27 years) and an average of 17.8 years of academic medicine experience (range 9–24 years). Their judgments were made while having access to all information types (C, I, and A) using the same displays as the resident participants in the CIA condition, with access to no additional information. Of note, two of these five attending physicians also participated in the focus group to design the algorithm to compute the aggregated quality score (A). The focus group was conducted five months prior to when the two attending physicians made their quality judgments as experts for our study.

In order to compare the resident physicians' judgments to the judgments made by the experts, the attending physicians' judgments for each of the 30 profiles were first averaged. Then, each resident participant's set of judgments was correlated with the set of the average expert's judgments to obtain a measure of accuracy of judgment. This measure is called "achievement" in JA studies. Single-factor ANOVA was then used to analyze the effect of information type on accuracy across the four information conditions.

To address the second objective of investigating the *consistency* of quality judgments, linear regression on the participants' (residents) judgments with the component quality measures was conducted. The coefficient of multiple correlation, called "cognitive control" in JA studies, was used as a measure of consistency of judgment. A participant with perfect consistency (i.e., the component measures had equal impact on the quality judgment for each profile) would have a consistency measure of 1.0 and would show that judgments were made in a consistent and controlled fashion over the 30 profiles. Again, single-factor

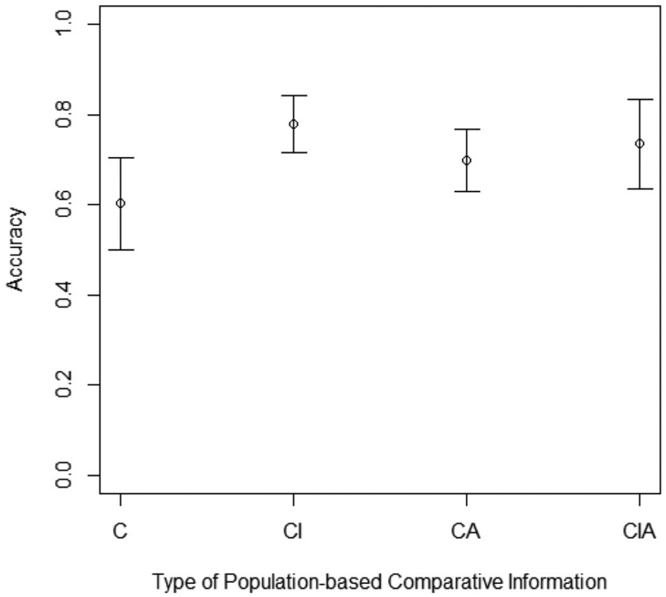


Fig. 2. Mean accuracy of participants' judgments by type of pooled comparative information with 95% confidence intervals ($n = 8$ resident physicians in each condition).

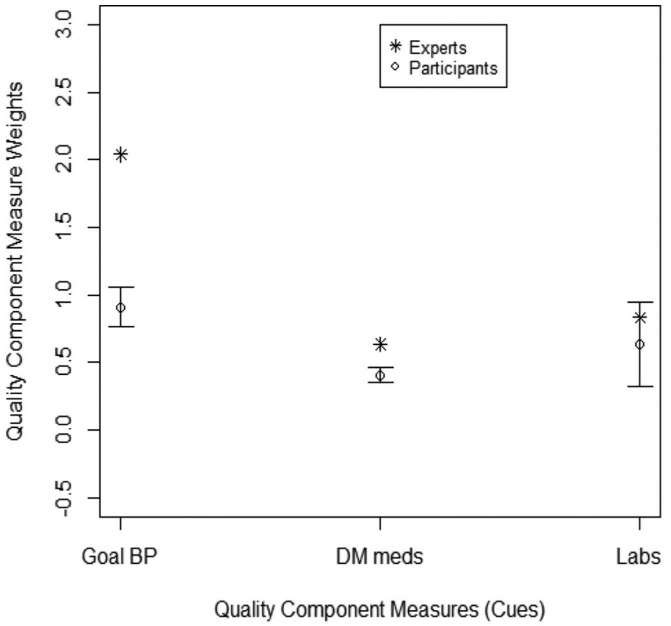


Fig. 3. Mean quality component measure (cue) weights for participants' and expert's models (95% confidence intervals shown for participants' weights).

ANOVA was used to analyze the effect of type of information on consistency across the four information conditions. For both the first and second objectives, Fisher's r to z_r transformation was performed on the correlations prior to conducting the ANOVA. Post-hoc analysis was also conducted using Tukey's Honestly Significant Difference.

To address the third objective of investigating the *reliability* of residents' quality judgments, the standard deviation in judgment values was calculated for each profile across participants grouped by experimental condition. Thus, for each experimental condition, there were 30 standard deviation values (one for each profile). A small standard deviation value would represent better reliability between residents for that profile. A repeated measures ANOVA was conducted with information condition nested within profile (i.e., to ensure the information condition effect within each and every profile). Post-hoc pairwise t-tests with adjusted p-values were also conducted.

III. Results

A. Impact of Pooled Comparative Information on Accuracy of Quality Judgment

Overall, the participants' ability to judge the quality of hypertension care in terms of accuracy was fair. Average accuracy measures (i.e., correlation of the participants' judgments with those made by experts) were 0.60, 0.78, 0.70, and 0.74 in the C, CI, CA, and CIA conditions, respectively. Fig. 2 shows the means of accuracy by information type. Single-factor ANOVA of transformed data indicated a significant effect of information type on accuracy: $F(3,28) = 4.31$; $p = 0.01$. Tukey's post-hoc analysis indicated that accuracy was significantly higher for the CI condition compared with the C condition ($p < 0.01$), and there was a trend for the CIA condition to have higher accuracy than the C condition ($p = 0.09$). There were no other significant differences between other conditions.

Although the expert judgments were significantly correlated with the composite quality of care score that was provided to residents in the CA and CIA conditions ($r = 0.92$, $p < 0.0001$), neither condition including that information resulted in higher resident correlation with the experts (to indicate higher accuracy in judgment). In fact, participants in the CI condition had the highest average correlation with the composite score compared with the other conditions; yet, these participants never saw the composite score. Average participant correlations with the composite score were 0.56, 0.78, 0.70, and 0.72 for C, CI, CA, and CIA, respectively.

The low accuracy across each experimental group may be attributed to mismatch in how the participants weighted the component measures as compared with the experts in making their quality judgments. Fig. 3 presents the computed measure weights for all participant models. The participants tended to underweight the measures as compared with the model from the experts. The resident participants' use of the patient outcome measure (percent of patients at goal BP) had less of an effect on judgment of quality of care compared with the experts. Considering the variance in the measure weights, the participants considered the recommended medications cue in a more consistent manner. Further, the process measure of percent of patients with recommended laboratory tests checked had a greater impact on five resident participants' judgments of quality of care compared with the experts. This could indicate that some residents may be more concerned with process measures, over which they have more control and may be less concerned with outcome measures that may be difficult to control in short periods of time (i.e., their residency program).

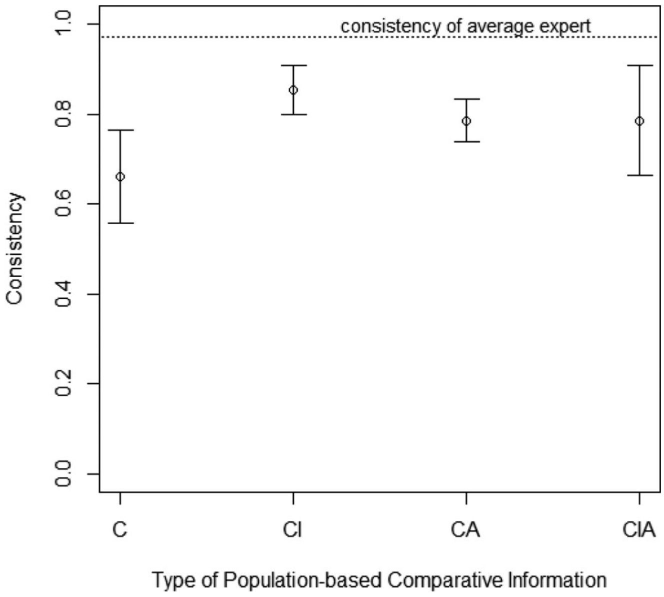


Fig. 4. Mean consistency of participants' judgments by type of pooled comparative information with 95% confidence intervals ($n = 8$ resident physicians in each condition). The consistency of the average expert (attending physicians) is indicated using the dotted line.

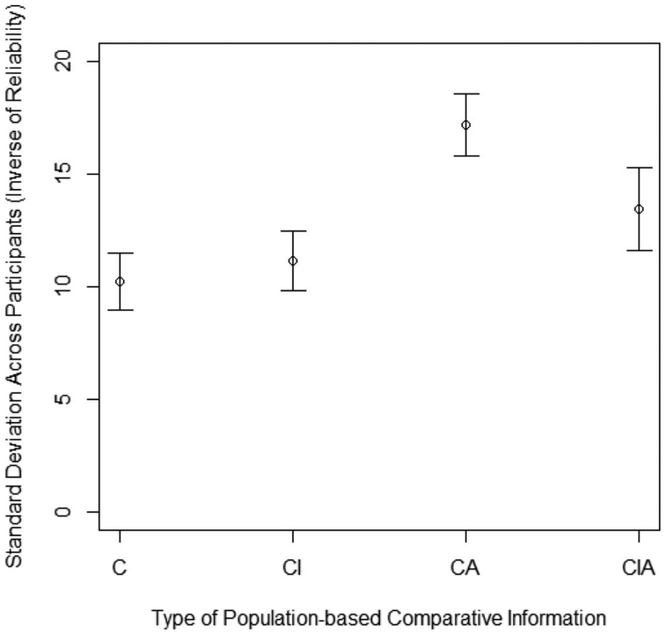


Fig. 5. Reliability as measured by (inverse of) mean standard deviations of judgments, grouped by type of pooled comparative information ($n = 30$ judgment profiles in each condition).

B. Impact of Pooled Comparative Information on Consistency of Quality Judgment

The coefficients of multiple correlation of the linear regression on the participants' judgments were used as measures of consistency within individual participants across their 30 quality judgments. The average consistency was 0.66, 0.85, 0.79, and 0.79 for the C, CI, CA, and CIA conditions, respectively. Fig. 4 depicts the means of consistency for each information condition. For comparison, the graph also includes a dotted line representing the consistency of the experts at 0.97. Single-factor ANOVA on transformed data indicated a significant effect of information type on consistency: $F(3,28) = 5.10$; $p = 0.006$. Tukey's post-hoc analysis indicated that consistency was significantly higher for the CI condition compared with the C condition ($p < 0.005$), and there was a trend for the CIA condition to be more consistent compared with the C condition ($p = 0.09$). There were no other significant differences between the other conditions.

C. Impact of Pooled Comparative Information on Reliability of Quality Judgments

To address the third objective of investigating reliability across participants as a function of type of information presented, the standard deviation of participants' judgments for each profile grouped by experimental condition was calculated. Smaller standard deviations would indicate better reliability across participants. The averages of standard deviations for the 30 profiles were 10.21, 11.16, 17.19, and 13.45 for the C, CI, CA, and CIA conditions, respectively. Fig. 5 shows the means of standard deviations by information condition. A repeated measures ANOVA with the information condition nested within profile shows that information type significantly impacted the reliability of the judgments across participants: $F(3,112) = 3.37$; $p = 0.02$. Post-hoc pairwise t-tests with adjusted p-values show that the CA condition resulted in significantly lower reliability (high standard deviation) compared with all other conditions: C ($p < 0.0001$), CI ($p < 0.0001$), and CIA ($p = 0.003$). Further, the CIA condition had significantly lower reliability (higher standard deviation) than the C ($p < 0.01$) and CI ($p < 0.05$) conditions.

IV. Discussion

A. Use of Comparative Information in Judging Quality

Judgments of quality based on pooled data are the focus of many quality programs and healthcare initiatives. Yet, it is not well understood how the type of information that is used in quality decision support tools impacts quality judgments. The goal of this study was to investigate the impact of different types of pooled comparative information on the ability of resident physicians' to judge quality of hypertension care.

In general, this was a difficult task for residents as indicated by somewhat low accuracy and internal consistency in each information condition. However, those who were presented with quality component measures and the addition of an interpretation of the component measures in the form of peer ranking (CI condition) had significantly more accurate judgments and were more internally consistent compared with residents presented with quality component measures alone. Accuracy was raised from 0.60 to 0.78, and consistency was raised from 0.66 to 0.85. Further, the residents in the CI condition were significantly more reliable between each other compared with those who were presented with an aggregation of component measures as a composite score (both the CA and CIA conditions). This was true despite little instruction, no training to a criterion, and no feedback on performance. This shows promise for this type of information to enhance the ability for one to make quality judgments. These results are supported by other studies of judgment, where additional contextual information improved judgment performance, particularly more than the addition of an algorithmic judgment [36]–[38].

Although participants in the CA and CIA condition were provided with a quality composite score (A) that closely replicated expert judgment, they did not achieve higher accuracy and consistency and were significantly less reliable between each other. This was a somewhat surprising result. It is possible that participants in those groups may have adopted a distrusting strategy that involved anchoring their judgment on the composite score and then moving away from that score [39]. Further, participants were not provided with information regarding how the algorithm aggregated the component measures to determine the composite score, which could have resulted in a lack of trust in the score. Other studies have shown that when judges are provided with explanations about an algorithmic judgment (such as its weighting strategy), performance can be improved [40]. In those cases, judges also tended to increase adaptation to the algorithm, even when unwarranted [41] and especially when their trust in the algorithm exceeded their self-confidence [42].

The presence of the aggregated quality score could also have negatively impacted performance due to participants employing strategies related to single-factor dominance theory [43]. Specifically, dissonance between the overall quality composite score and the quality component measure of percentage of patients at goal blood pressure (deemed most important by the experts) may have caused the participants to distrust the algorithm. Participants may have biased their judgments away from the algorithm's score (and further from the experts) when they perceived this dissonance.

B. Limitations and Future Work

There is a broader range of types and display representations of comparative information we could have presented. In particular, our representation for quality component measures (C) in the form of dot plots showing means and variations did promote some interpretation of the information. Additional representations of the types of comparative information should be tested to further understand the impact on quality judgment. For example, performance polygons have been suggested as another way to present component quality measures [44]. We also only provided two filters to interrogate the component measures when judging quality of hypertension care (diabetes and low income patients only). It is possible that providing additional filtering functionality could have improved the residents' judgments. Further, there may be a broader set of quality component measures that could be used to make a judgment regarding the quality of hypertension care and alternate methods for measure identification may be employed. For example, Cooksey [25] also suggests using a verbal protocol analysis for cue (component measure) identification.

The algorithm used to aggregate the component measures into a composite quality score was derived using expert opinion (attending physicians). The algorithm was highly correlated with the experts, whose quality judgments were used as the gold standard for comparison, implying that it was a realistic model of expert judgment. However, we do know there was a mismatch in the weighting scheme used to aggregate the component measures between the participants (here the residents) and the experts (here attending physicians). Others have found success in the use of composite scorecards when involving users in all aspects of measure selection and weighting [45]. Future work should investigate the impact of aggregating component measures based on the participant models to see if this impacts accuracy, consistency, and reliability of judgment.

Another limitation to this study was that experts' judgments were used as the gold standard for quality of care judgments. Their judgments could have been biased to the focus of the clinic or to the individual pressure or interest they each have related to hypertension care. For example, it could be that expert attending physicians are more interested in maintaining goal blood pressure among their hypertensive patients because that outcome is under more scrutiny by insurance companies or hospital administrators. Despite this limitation, the JA approach enabled uncovering the impact of information type on both consistency within individual participants (residents) and reliability across participants, which were analyzed independent of the experts' judgments.

The quality component measures used in this study were obtained from a clinical data repository containing only EMR-derived data. Database integrity is, therefore, critical to the success of the quality judgment program. Some studies have shown that quality component measures may not represent the true population (here, patients seeking care) due to numerator loss and thus underestimation of process measures [13]–[15]. When developing quality component measures and related information to use in quality judgment, one should consider the limitations of the available data. Another data limitation in this study was the use of confidence intervals derived after scaling the comparative populations, which reduced the size of the intervals. This design choice was made to limit the possibility of confounding the experiment if some users disregarded the data altogether and were unable to make a judgment based on comparing populations. In general, the use of confidence intervals for smaller populations is a major challenge in communicating uncertainty. Future work should investigate alternate strategies for displaying uncertainty in the component measures that could enable better judgment performance. Potential options could be the use of shading [46] or textures, or using a variant of the box plot to indicate descriptive data features [46]. Natural language processing techniques may also help improve the validity of quality measures derived from unstructured data sources, reducing numerator loss and improving the usefulness of confidence intervals [47].

From an application perspective, it is important to understand the goal of quality decision support tools and the impact that the design of such tools have on quality judgment. It is unclear if clinicians who are better able to judge the quality of their care (i.e., make more accurate, consistent, and reliable judgments) based on pooled comparative information will improve their practice behaviors, resulting in better care for the patients. Additionally, there could be other unintended consequences of poor quality judgment performance. For example, physicians may avoid certain populations of patients or discount clinical knowledge and patient preference in order to achieve better quality component measures that are weighted more heavily [48]. Future work should ultimately aim to investigate the impact of the quality assessment process on specific practice behaviors and patient care.

V. Conclusion

Judging quality, especially in healthcare, is difficult. We need decision support tools that support accurate, consistent, and reliable judgments of quality in order to enhance patient care [49]. True quality judgments allow practitioners to investigate and evaluate their care in order to recognize successes and identify areas for improvement [45]. The results of this study have implications for the design of quality decision support tools. For such tools that aim to support one's ability to judge quality based on pooled comparative information, it may be more beneficial to focus design efforts on aiding the interpretation of quality measures, rather than on developing sophisticated algorithms that aggregate quality measures into composite scores.

Acknowledgment

The authors would like to acknowledge the internal medicine attending and resident physicians at the University of Virginia, without whom this work could not have been completed. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine, the National Institutes of Health, or the authors' institutions. None of these organizations had any influence in the design and conduct of the study, the collection, management, analysis, and interpretation of the data; or preparation, review, or approval of this manuscript.

References

[1] U.S. Congress, *Medicare Improvements for Patients and Providers Act of 2008*, 2008.

[2] U.S. Congress, *Patient Protection and Affordable Care Act of 2010*, 2010.

[3] P. D. Tucker, *Linking Teacher Evaluation and Student Learning*. Alexandria, VA, USA: Assoc. Supervision Curriculum Develop., 2005.

[4] American Board of Medical Specialties. (2014). *ABMS Maintenance of Certification*. [Online]. Available: http://www.abms.org/Maintenance_of_Certification/ABMS_MOC.aspx

[5] Accreditation Council of Graduate Medical Education. (2013). *The ACGME Outcome Project*. [Online]. Available: http://www.acgme.org/acgmeweb/Portals/0/PFAssets/ProgramRequirements/CPRs2013.pdf

[6] D. A. Davis, P. E. Mazmanian, M. Fordis, R. Van Harrison, K. E. Thorpe, and L. Perrier, "Accuracy of physician self-assessment compared with observed measures of competence: A systematic review," *J. Am. Med. Assoc.*, vol. 296, no. 9, pp. 1094–1102, 2006.

[7] (2014). *HEDIS 2015 Summary Table of Measures, Product Lines and Changes*. [Online]. Available: http://www.ncqa.org/Portals/0/HEDISQM/Hedis2015/List_of_HEDIS_2015_Measures.pdf

[8] S. M. Campbell, "Research methods used in developing and applying quality indicators in primary care," *Quality Safety Health Care*, vol. 11, no. 4, pp. 358–364, Dec. 2002.

[9] *Physicians' Views of Comparative Information on Costs and Resource Use: Findings and Implications for Report Developers*, Robert Wood Johnson Foundation, Princeton, NJ, USA, 2012.

[10] M. B. Rothberg, E. Morsi, E. M. Benjamin, P. S. Pekow, and P. K. Lindenauer, "Choosing the best hospital: The limitations of public quality reporting," *Health Aff. (Millwood)*, vol. 27, no. 6, pp. 1680–1687, Nov. 2008.

[11] P. Barkhuysen, W. de Grauw, R. Akkermans, J. Donkers, H. Schers, and M. Biermans, "Is the quality of data in an electronic medical record sufficient for assessing the quality of primary care?" *J. Am. Med. Inf. Assoc.*, vol. 21, no. 4, pp. 692–698, Aug. 2014.

[12] G. Sidorenkov, F. M. Haaijer-Ruskamp, D. de Zeeuw, H. Bilo, and P. Denig, "Relation between quality-of-care indicators for diabetes and patient outcomes: A systematic literature review," *Med. Care Res. Rev.*, vol. 68, no. 3, pp. 263–289, May 2011.

[13] A. Parsons, C. McCullough, J. Wang, and S. Shih, "Validity of electronic health record-derived quality measurement for performance monitoring," *J. Am. Med. Inf. Assoc.*, vol. 19, no. 4, pp. 604–609, Jan. 2012.

[14] K. S. Chan, J. B. Fowles, and J. P. Weiner, "Electronic health records and the reliability and validity of quality measures: A review of the literature," *Med. Care Res. Rev.*, vol. 67, no. 5, pp. 503–527, Feb. 2010.

[15] K. Dentler, R. Cornet, A. Teije, P. Tanis, J. Klinkenbijl, K. Tytgat, and N. Keizer, "Influence of data quality on computed Dutch hospital quality indicators: A case study in colorectal cancer surgery," *BMC Med. Inf. Decis. Mak.*, vol. 14, no. 1, p. 32, 2014.

[16] Centers for Medicare & Medicaid Services. *Hospital Compare*. [Online]. Available: http://www.medicare.gov/hospitalcompare/search.html

[17] The Leapfrog Group. (2015). *Leapfrog Hospital Survey Results*. [Online]. Available: http://www.leapfroggroup.org/cp

[18] C. Schoen, K. Davis, S. K. H. How, and S. C. Schoenbaum, "U.S. Health System Performance: A national scorecard," *Health Aff. (Millwood)*, vol. 25, no. 6, pp. w457–w475, Nov. 2006.

[19] A. M. Zaslavsky, J. A. Shaul, L. B. Zaborski, M. J. Cioffi, and P. D. Cleary, "Combining health plan performance indicators into simpler composite measures," *Health Care Finance Rev.*, vol. 23, no. 4, pp. 101–115, 2002.

[20] E. Petra, P. Varughese, L. Epifania, L. Buneo, and K. Scarfone, "Use of quality index tracking to drive improvement in clinical outcomes," *Nephrol. News Issues*, vol. 20, no. 8, pp. 67–68, Jul. 2006.

[21] J. Profit, K. V. Typpo, S. J. Hysong, L. D. Woodard, M. A. Kallen, and L. A. Petersen, "Improving benchmarking by using an explicit framework for the development of composite indicators: An example using pediatric quality of care," *Implement. Sci.*, vol. 5, no. 1, p. 13, 2010.

[22] E. S. Holmboe, W. Weng, G. K. Arnold, S. H. Kaplan, S.-L. Normand, S. Greenfield, S. Hood, and R. S. Lipner, "The Comprehensive Care Project: Measuring physician performance in ambulatory practice," *Health Serv. Res.*, vol. 45, no. 6, p. 2, pp. 1912–1933, Dec. 2010.

[23] A. D. Simms, P. D. Baxter, B. A. Cattle, P. D. Batin, J. I. Wilson, R. M. West, A. S. Hall, C. F. Weston, J. E. Deanfield, K. A. Fox, and C. P. Gale, "An assessment of composite measures of hospital performance and associated mortality for patients with acute myocardial infarction. analysis of individual hospital performance and outcome for the national institute for cardiovascular outcomes research (NICOR)," *Eur. Heart J. Acute Cardiovasc. Care*, vol. 2, no. 1, pp. 9–18, Mar. 2013.

[24] R. Jacobs, M. Goddard, and P. C. Smith, "How robust are hospital ranks based on composite performance measures?," *Med. Care*, vol. 43, no. 12, pp. 1177–1184, Dec. 2005.

[25] R. Cooksey, *Judgment Analysis: Theory, Methods, and Applications*. San Diego, CA, USA: Academic, 1996.

[26] C. J. Hursch, K. R. Hammond, and J. L. Hursch, "Some methodological considerations in multiple-cue probability studies," *Psychol. Rev.*, vol. 71, no. 1, pp. 42–60, 1964.

[27] L. R. Tucker, "A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd," *Psychol. Rev.*, vol. 71, no. 6, pp. 528–530, 1964.

[28] A. D. MacCormick and B. R. Parry, "Judgment analysis of surgeons' prioritization of patients for elective general surgery," *Med. Decis. Making*, vol. 26, no. 3, pp. 255–264, 2006.

[29] A. T. Hirsh, S. B. Callander, and M. E. Robinson, "Patient demographic characteristics and facial expressions influence nurses' assessment of mood in the context of pain: A virtual human and lens model investigation," *Int. J. Nurs. Stud.*, vol. 48, no. 11, pp. 1330–1338, Nov. 2011.

[30] A. V. Chobanian, G. L. Bakris, H. R. Black, W. C. Cushman, L. A. Green, J. L. Izzo, Jr., D. W. Jones, B. J. Materson, S. Oparil, J. T. Wright, Jr., and E. J. Roccella, "The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: The JNC 7 report," *J. Am. Med. Assoc.*, vol. 289, no. 19, pp. 2560–2572, May 2003.

[31] *National Voluntary Consensus Standards for Ambulatory Care Using Clinically Enriched Administrative Data: A Consensus Report*, National Quality Forum, Washington, DC, USA, 2010.

[32] Centers for Medicare and Medicaid Services. (2011, Apr. 20). *Physician Quality Reporting System formerly known as the Physician Quality Reporting Initiative*. [Online]. Available: http://www.cms.gov/PQRS//

[33] Agency for Healthcare Research and Quality. (2011). *National Quality Measures Clearing House*. [Online]. Available: http://www.qualitymeasures.ahrq.gov/

[34] National Committee for Quality Assurance. (2011). *The Healthcare Effectiveness Data and Information Set (HEDIS)*. [Online]. Available: http://www.ncqa.org/HEDISQualityMeasurement/HEDISMeasures.aspx

[35] L. A. Baumgart, E. J. Bass, J. A. Lyman, S. Springs, J. Voss, G. F. Hayden, M. A. Hellems, T. R. Hoke, K. A. Schlag, and J. B. Schorling, "Supporting physicians' practice-based learning and improvement (PBLI) and quality improvement through exploration of population-based medical data," in *Proc. 54th Annu. Meet. Human Factors Ergonomic Soc.*, 2010, vol. 54, pp. 845–849.

[36] E. J. Bass, L. A. Baumgart, and K. K. Shepley, "The effect of information analysis automation display content on human judgment performance in noisy environments," *J. Cogn. Eng. Decis. Making*, vol. 7, no. 1, pp. 49–65, Aug. 2012.

[37] G. Gattie and A. Bisantz, "The effects of integrated cognitive feedback components and task conditions on training in a dental diagnosis task," *Int. J. Ind. Ergon.*, vol. 36, no. 5, pp. 485–497, 2006.

[38] F. A. Drews and D. R. Westenskow, "The right picture is worth a thousand numbers: Data displays in anesthesia," *Human Factors*, vol. 48, no. 1, pp. 59–71, 2006.

[39] E. J. Bass and A. R. Pritchett, "Human-Automated Judge Learning: A methodology for examining human interaction with information analysis automation," *IEEE Trans. Syst. Man Cybern. A., Syst. Humans*, vol. 38, no. 4, pp. 759–776, Jul. 2008.

[40] Y. Seong and A. Bisantz, "The impact of cognitive feedback on judgment performance and trust with decision aids," *Int. J. Ind. Ergon.*, vol. 38, nos. 7/8, pp. 608–625, 2008.

[41] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Human-Comput. Stud.*, vol. 58, no. 6, pp. 697–718, 2003.

[42] J. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *Int. J. Human-Comput. Stud.*, vol. 40, no. 1, pp. 153–184, 1994.

[43] H. Montgomery and O. Svenson, "A think aloud study of dominance structuring in decision processes," in *Process and Structure in Human Decision Making*. Chichester, U.K.: Wiley, 1989, pp. 135–150.

[44] T. M. Cook, M. Coupe, and T. Ku, "Shaping quality: The use of performance polygons for multidimensional presentation and interpretation of qualitative performance data," *Brit. J. Anesth.*, vol. 108, no. 6, pp. 953–960, Mar. 2012.

[45] S. A. Fields and D. Cohen, "Performance enhancement using a balanced scorecard in a patient-centered medical home," *Family Med.*, vol. 43, no. 10, pp. 735–739, Dec. 2011.

[46] A. M. Bisantz, D. Cao, M. Jenkins, P. R. Pennathur, M. Farry, E. Roth, S. S. Potter, and J. Pfautz, "Comparing uncertainty visualizations for a dynamic decision-making task," *J. Cogn. Eng. Decis. Making*, vol. 5, no. 3, pp. 277–293, Sep. 2011.

[47] K. Dentler, M. E. Numans, A. ten Teije, R. Cornet, and N. F. de Keizer, "Formalization and computation of quality measures based on electronic medical records," *J. Am. Med. Inf. Assoc.*, vol. 21, pp. 285–291, Nov. 2013.

[48] R. M. Werner, "The unintended consequences of publicly reporting quality information," *J. Am. Med. Assoc.*, vol. 293, no. 10, pp. 1239–1244, Mar. 2005.

[49] S. Stone-Griffith, J. D. Englebright, D. Cheung, K. M. Korwek, and J. B. Perlin, "Data-driven process and operational improvement in the emergency department: The ED Dashboard and Reporting Application," *J. Healthcare Manag.*, vol. 57, no. 3, pp. 167–180, Jun. 2012.