Assessment of the Scientific Soundness of Clinical Performance Measures

A Field Test of the National Committee for Quality Assurance’s Colorectal Cancer Screening Measure

Eric C. Schneider, MD; Marion R. Nadel, PhD; Alan M. Zaslavsky, PhD; Elizabeth A. McGlynn, PhD

Background: Relatively few studies have evaluated the scientific soundness of widely used performance measures. This study evaluated quality measures by describing a field test of the colorectal cancer screening measure included in the Health Plan Employer Data and Information Set of the National Committee for Quality Assurance.

Methods: We conducted a field test in 5 health care plans that enrolled 189,193 individuals considered eligible for colorectal cancer screening. We assessed measurement bias by calculating the prevalence of colorectal cancer screening while varying the data sources used (administrative data only, a hybrid of administrative data and medical record data, and enrollee survey data only) and the minimum required enrollment period (2-10 years).

Results: Across the 5 health care plans, the percentage of health care plan enrollees counted as screened varied according to the data used, ranging from 27.3% to 47.1% with the administrative data, 38.6% to 53.5% with the hybrid data, and 53.2% to 69.7% with the survey data. The relative ranking of plans also varied. One health care plan ranked first based on administrative data, second based on hybrid data, and fourth based on survey data. Survey respondents were more likely than nonrespondents to have evidence of colorectal cancer screening (62.7% vs 46.5%; P < .001).

Conclusions: Administrative data seem to underestimate colorectal cancer screening and survey data seem to overestimate it, suggesting that a hybrid data approach offers the most accurate measure of screening. Implementation of performance measures should include evaluation of their scientific soundness.

Arch Intern Med. 2008;168(8):876-882

 Efforts to measure and report on the quality of health care have expanded rapidly in the United States.1-3 Many performance measures used to assess health care plans are now used to assess the ambulatory care performance of physician groups. Higher performance on process measures seems to be associated with better health outcomes, and performance reporting may lead to improvements in the quality of clinical care.4-9 Nevertheless, performance measurement in health care remains controversial, in part because of concerns that measures with poor validity will drive physicians to provide inappropriate care, even when incentives are attached to the measures.10,11

The validity of clinical performance measures is determined largely by measure specifications and the availability of detailed clinical data.14 However, relatively few evaluations of the scientific soundness of measure specifications have been published.15-20 Developing and implementing valid performance measurement can be costly. To minimize data collection costs, measure developers often rely solely on electronic administrative, claims, or registry data without formally evaluating the impact of these choices on measure validity.

In 2004, the National Committee for Quality Assurance (NCQA) introduced a clinical performance measure of colorectal cancer screening prevalence into the Health Plan Employer Data and Information Set (HEDIS), a nationally standardized widely used set of clinical performance measures.21 A high level of scientific evidence supports the effectiveness of colorectal cancer screening for specific populations, and screening is endorsed by all major US guidelines, but rates of screening remain low.22-26

Colorectal cancer screening poses 3 key measurement challenges. First, unlike other clinical preventive services (eg, mammography) that consist of a single test per-
formed at regular intervals, colorectal cancer screening can involve any of the following 4 screening tests: annual fecal occult blood testing (FOBT) with a home test kit, sigmoidoscopy every 5 years, colonoscopy every 10 years, and air- or double-contrast barium enema every 5 years. Major national guidelines do not yet favor any particular test, although specialty society guidelines express a preference for colonoscopy.27,28 Second, the 4 guideline-defined screening strategies involve screening intervals that range from 1 to 10 years, with longer intervals potentially challenging the capability of records and patient recall. Third, the choice of screening strategy depends on detailed clinical data about the patient's risk for colorectal cancer. This article describes a field test that assesses the potential for measurement bias related to each of these issues.

**METHODS**

**OVERVIEW**

Performance measure specifications define the sampling of enrollees and the data sources used (Table 1).29 Bias can arise in many ways. For example, until recently, the Current Procedural Terminology (CPT) code for FOBT could apply to either the FOBT home test kit (an acceptable test for colorectal screening) or the single in-office test (not an acceptable test for colorectal screening).30 In the absence of a gold standard, measure evaluators can assess the convergent validity of alternative measure specifications, comparing results obtained using alternative specifications of data sources and combinations of data used, the time window in which care is assessed, and inclusion and exclusion criteria. Convergent validity can be assessed using different specifications within the same sample of enrollees or using alternative specification-defined samples from the same health care plans (assuming selection is random).

The term administrative data refers to the health care plan files that contain a combination of beneficiary characteristics (eg, age, sex, race/ethnicity, and residence) and claims data (the service and diagnostic codes submitted by physicians). In addition, the NCQA has developed a hybrid method that starts with a search of administrative data for evidence that a clinical service was provided and is followed by a review of the medical records of patients who seem, based on administrative data, not to have received the service.31 This approach limits the number of required medical record reviews and may compensate for data not contained in administrative files. Enrollee surveys, such as the Consumer Assessment of Healthcare Providers and Systems survey, have been used to assess services such as influenza vaccination that can be delivered in nonclinical settings and not recorded in administrative data or medical records.32 The measurement time window is another potential source of bias. HEDIS measures specify a minimum required continuous enrollment period—the interval during which a person must be enrolled in a health care plan without a break in insurance. For most screening measures, the continuous enrollment period is set equal to the interval for subsequent screening. A longer minimum enrollment requirement reduces the sample size for

---

**Table 1. Comparison of Strengths and Weaknesses of Data Collection Methods for Performance Measurement**

<table>
<thead>
<tr>
<th>Data Collection Method</th>
<th>Basic Approach</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative</td>
<td>Identify all persons who meet the eligibility criteria; among those, use standard coding methods (eg, CPT 4) to determine whether those persons were screened during the appropriate interval</td>
<td>Method can be applied to the entire enrolled population; relatively inexpensive after initial investment in developing algorithms; may facilitate quality improvement interventions (by identifying specific individuals who have not been screened)</td>
<td>Variability in coding practices can reduce comparability of results across plans; some reimbursement methods or plan types may not track some services or tests; information only available for the time an individual was enrolled in the plan; may not be able to distinguish between screening asymptomatic persons and evaluating symptoms</td>
</tr>
<tr>
<td>Medical record abstraction</td>
<td>From administrative data, select a random sample of eligible persons; using a standard form, abstract data from medical records of ≥1 physicians to determine if screening was received</td>
<td>Clinically detailed (exclusions, risk factors, and patient refusal can be identified); availability of data not linked to a particular reimbursement method; historical information available</td>
<td>Expensive and time-consuming; may have to review multiple records; recordkeeping practices not fully standardized across physicians or health care plans; privacy concerns may limit ability to use this method; imposes a burden on physicians; not easily used for targeted quality improvement efforts</td>
</tr>
<tr>
<td>Hybrid method (administrative plus medical record abstraction)</td>
<td>Apply the administrative method (detailed previously); select a random sample of persons who seem, based on administrative data, not to have received screening; apply the medical record abstraction method to this sample to determine if screening was received</td>
<td>Corrects for bias when administrative data are incomplete; minimizes the required number of medical record reviews; partially addresses weaknesses of the administrative and medical record abstraction methods; may facilitate improvement in coding of administrative claims</td>
<td>Sampling strategy does not enable estimation of the proportion of administrative false positives; medical record abstractions unnecessary if administrative data needed for the measure are complete and accurate</td>
</tr>
<tr>
<td>Enrollee survey</td>
<td>From administrative data, select a random sample of eligible persons; obtain information using a standard mail or telephone questionnaire</td>
<td>Enrollee can report events aggregated across insurance and physician types; may allow exploration of reasons for nonadherence to clinical recommendations</td>
<td>Nonresponse can create biased results; low reliability of self-report for long periods; expensive; surveys are usually anonymous, limiting utility of the data for quality improvement</td>
</tr>
</tbody>
</table>

statistical comparisons and biases the sample toward long-term enrollees, who may differ systematically from others. Misclassification within the denominator is another source of bias. High-risk enrollees may require more frequent screening than the measure specifies, but administrative data may not identify risk factors such as a family history of colorectal cancer or polyps or a personal history of polyps. Inclusion of high-risk enrollees can lead plans that enroll more of them to spuriously seem more successful at screening for average-risk enrollees.

APPROACH TO THE FIELD TEST

Data Sources

We defined, for each data source (administrative, medical record, and survey results), the data elements needed to implement all of the measure specifications. We developed a list of outpatient diagnosis codes that represent a prior diagnosis of colorectal cancer, CPT codes related to an acceptable screening procedure, and historical CPT codes used within the previous 10 years. For the medical record method, we designed an abstraction tool to collect data on screening tests, clinical risk factors for colorectal cancer, and evidence of limited life expectancy and training experienced nurses from each health care plan during a pilot test of the abstraction protocol using a common set of records. For the enrollee survey, we developed a sequence of questions (based on the Behavioral Risk Factor Surveillance System) that addressed each of the screening tests, the time frames in which they occurred, and a measure of risk status based on report of a family history of colorectal cancer.

Data Collection

The NCQA recruited 5 geographically dispersed health care plans that represent a variety of organizational types to participate in the field test. Each health care plan identified all enrollees 51 years or older as of December 31, 1999, who had been continuously enrolled in the health care plan for at least 2 years (January 1, 1998, through December 31, 1999). Health care plan analysts extracted the entire administrative and claims data history and sent it to the RAND Corporation for further sampling and analysis (Figure). To implement medical record abstraction for the NCQA hybrid method, research staff at the RAND Corporation selected, at random, 1000 enrollees (200 per plan) who lacked evidence of colorectal cancer screening based on administrative data. Medical record abstractors at each plan abstracted data from these enrollees’ primary care records. The abstractions were returned to the RAND Corporation for data entry and analysis. Each health care plan’s survey sample consisted of the 200 enrollees whose records were abstracted plus an additional random sample of 400 enrollees from the original health care plan cohort (3000 enrollees). The survey was administered via mail by the RAND Corporation’s Survey Research Group. During a 6-week period, nonrespondents received 3 survey mailings plus a reminder postcard and a final overnight mailing. The response rate was 48.1% and varied across health care plans from 37.8% to 57.5%.

Analysis

Data from all sources (administrative, survey, and medical record) were linked at the enrollee level to create a single analytic file. We obtained data on the region, model type, age, and total enrollment of the 5 health care plans from InterStudy data (based on reports by the plans to an annual survey). We calculated the total number of sampled members, their mean age, the proportion of females, and the mean length of enrollment. We tested differences among plans using a chi-square test for sex and analysis of variance for age and length of enrollment. For each of the 4 colorectal screening tests, we compared across health care plans the percentage of sampled enrollees identified as screened based on administrative data and the percentage based on survey data. We could not determine the percentage screened by each test under the hybrid method because medical record abstractors were instructed to stop after identifying the occurrence of 1 of the tests. For the single measure that combined all tests, we calculated colorectal cancer screening performance scores based on administrative data only, survey data only, and combined administrative and medical record data (the hybrid method), weighing the medical record sample to represent the population from which they were drawn (unscreened based on administrative data). We calculated 95% confidence intervals for each estimate. Among the survey respondents, we assessed agreement of the sampled enrollees’ screening status according to the survey data and the hybrid data. We compared the rate of screening among survey respondents and nonrespondents using the hybrid estimation procedure. Finally, across the 4 health care plans that supplied enrollment dates, we compared screening rates, plan rankings, and the number of enrollees in the denominator for varied lengths of continuous enrollment (2, 3, 4, 5, 7, and 10 years).

Plans varied on total enrollment (from approximately 114 000 to 650 000), model type, and the number of years in operation (Table 2). Beneficiaries’ mean age, sex, and mean length of enrollment also differed significantly across the health care plans. Compared with survey nonrespondents, respondents were older (60.4 vs 59.4 years; P < .001) and had longer enrollment (73.3 vs 67.6 months; P = .001), but the 2 groups had similar percentages of female participants (53.3% vs 51.1%; P = .28). The percentage of enrollees having specific tests varied between the administrative and survey data methods (Table 3). In health care plans A and E, the percentages of enrollees screened by FOBT according to the 2 methods were similar; in health care plan B, the percentage based on survey data was nearly twice that based on administrative
estimates based on the hybrid and survey methods was mod-
the other 3 plans had fairly similar rankings.
Comparing the hybrid and survey methods, the relative rank-
on hybrid data, and ranked fourth based on survey data.

Using the single measure of screening (based on any
of colorectal cancer screening varied substan-
tially depending on the data sources used (Table 4).

Using the administrative data, the calculated percent-
age of eligible enrollees screened ranged from 27.3% (health
care plan D, which lacked FOBT claims) to 47.1% (health
care plan C); using the hybrid method, the cal-
culated percentage ranged from 38.6% (health care plan D)
to 53.5% (health care plan B); and using the survey
method, the percentage ranged from 53.2% (health care
plan A) to 69.7% (health care plan B). For 4 of the 5 health
care plans, the hybrid method produced a higher screen-
ing rate than the administrative data method. The dif-
cference ranged from 0 (health care plan A) to 14.9 (health
care plan B) percentage points. For all 5 health care plans,
the survey data method produced a higher calculated screening rate than the other 2 methods. The relative rank-
ing of the 5 health care plans varied according to the data
source used. For example, health care plan C ranked high-
est based on administrative data, ranked second based
on hybrid data, and ranked fourth based on survey data.
Comparing the hybrid and survey methods, the relative rank-
ing of plans C and D switched by 2 positions, but the other 3 plans had fairly similar rankings.

The level of agreement between the screening rate es-
timates based on the hybrid and survey methods was mod-
est (65.4%; k=0.34; P < .001). Among survey respondents
who reported having been screened within the appropriate interval, nearly half had no evidence of screen-
ing based on the hybrid method. In contrast, among the
38.1% of health care plan members with evidence of screening based on the hybrid method, only 13.9% re-
ported by survey that they had not received screening.

Using the hybrid data, survey respondents were more
likely than nonrespondents to have evidence of colorec-
tal cancer screening (62.7% vs 46.5%; P < .001). Among
the survey respondents, 15% reported a family history of colorectal cancer in a first-degree relative.
Lengthening the minimum required enrollment period reduced the number of plan members eligible for inclusion in the measurement sample (Table 5). The extent of the decline varied among the 4 health care plans that provided dates of enrollment. All 4 health care plans generated an adequate sample of eligible enrollees when the minimum required enrollment was less than 7 years. Three health care plans (C, D, and E) had few sampled members if the minimum enrollment period was set at 10 years. According to results based on administrative data, extension of the enrollment period beyond 2 years had a negligible effect on the percentage of plan members who appeared to be screened and no effect on the ranking of the health care plans relative to one another (data not shown). Among the patients sampled for medical record review, none had limited life expectancy. For all of the 5 health care plans we studied. Using the hybrid method as a comparison, plan-to-plan variations in the availability of electronic CPT codes for tests such as FOBT and flexible sigmoidoscopy in administrative data seem to bias comparisons of colorectal cancer screening rates. Our results are consistent with prior studies suggesting that administrative data may suffice for evaluating some aspects of health care (such as surgical procedure or mortality rates) but may not be suitable for evaluating others. Our results extend this literature by revealing the extent of plan-to-plan variation in the availability of specific administrative data elements. Estimates of screening rates based on survey data seemed to be biased by survey nonresponse, with nonrespondents less likely than respondents to have received screening for colorectal cancer. An unexpectedly high percentage of survey respondents reported a family history of colorectal cancer in a first-degree relative (15%) compared with a population-based sample of National Health Interview Survey participants (8%) and a sample from a community-based general medical practice (3%). Estimates of colorectal cancer screening prevalence are often drawn from self-reported survey data.

Our study illustrates potential threats to the validity of a clinical performance measure and highlights the value of field testing. The choice of data source had a substantial effect on the calculated colorectal cancer screening rates for all of the 5 health care plans we studied. Using the hybrid method as a comparison, plan-to-plan variations in the availability of electronic CPT codes for tests such as FOBT and flexible sigmoidoscopy in administrative data seem to bias comparisons of colorectal cancer screening rates. Our results are consistent with prior studies suggesting that administrative data may suffice for evaluating some aspects of health care (such as surgical procedure or mortality rates) but may not be suitable for evaluating others. Our results extend this literature by revealing the extent of plan-to-plan variation in the availability of specific administrative data elements.
despite known variation in the sensitivity and specificity of such self-reports.\textsuperscript{41,43} Like most health care plan surveys, ours involved a mailed questionnaire, a method that usually yields low response rates, heightening the potential for bias. The length of the continuous enrollment requirement had little effect on colorectal cancer screening rates and did not alter the ranking of health care plans relative to one another, suggesting that plan-related differences in enrollee turnover rates and the availability of data beyond a 5-year “look-back” period have little influence on the measure results. With the exception of the 10-year requirement, all of the study health care plans produced adequate sample sizes. This finding supports the use of the 2-year continuous enrollment requirement, which maximizes the available sample size without introducing bias.

Our study has limitations. We lacked a gold standard measure of colorectal screening. We had extensive data but a small sample of health care plans. The medical record abstraction protocol did not obtain the enrollee’s entire screening history. Our study did not address all of the potential threats to the validity of the colorectal cancer screening measure. We did not vary the age limit for screening eligibility (we assumed an age limit of 80 years) or test alternate clinical exclusion criteria, such as comorbid conditions or the full range of specific risk factors that would indicate a need for more frequent screening.

Limitations of the NCQA hybrid measure of colorectal screening also deserve mention. First, when the measure was introduced, CPT coding did not distinguish between the FOBT home test and the single in-office test (which is not an acceptable screen for colorectal cancer).\textsuperscript{30} A new CPT code specific to the home test kit for FOBT has been developed and measure specifications will incorporate this new code. Second, the measure specifications assume that screening for polyps and cancer inevitably occurs during any sigmoidoscopy, colonoscopy, or barium enema procedure performed for diagnostic purposes. If a significant proportion of these diagnostic tests are not adequate as screening, then the results may overstate the screening rate. Third, the measure specifications do not reflect the distinct screening requirements of health care plan members with a higher-than-average risk of colon cancer. For example, patients with a familial colon polyp syndrome might require a colonoscopy as frequently as every 3 years, but such patients are erroneously counted as screened if colonoscopy was performed within the previous 10 years. Fourth, the measure specifications do not encompass all of the causes of limited life expectancy that may justify forgoing colorectal cancer screening. Notably, our medical record abstraction found no enrollee with evidence of limited life expectancy. Measures typically cannot anticipate all of the potential exclusion criteria, so perfect performance scores might suggest inappropriate screening of some patients.

Given current data sources, the HEDIS hybrid measure of colorectal cancer screening rates seems valid for comparing the performance of health care plans. The hybrid method seems less prone to the between-plan bias associated with limitations of administrative data and the nonresponse bias associated with patient survey. The results of this field test led the NCQA to adopt the hybrid data method (combining administrative and medical record data) and to set a 2-year minimum continuous enrollment period for inclusion of enrollees in the measure of screening rates. Further improvement to specifications may be possible only as better clinical data become available.\textsuperscript{44}

Performance measures clearly have the potential to affect clinical practice (for better and for worse), especially when results are released to the public or used to set financial incentives for physicians.\textsuperscript{44,45} Our study illustrates a convergent validity method for measure evaluation in the absence of a gold standard of measurement. The Institute of Medicine has called for greater transparency of the US health care system. Field testing of measures should be a cornerstone of this transparency.\textsuperscript{16-20} Lacking rigorous evaluation, performance measures are unlikely to gain the credibility necessary to support improvements in health care quality.

Accepted for Publication: November 1, 2007.

Correspondence: Eric C. Schneider, MD, Department of Health Policy and Management, Harvard School of Public Health, 677 Huntington Ave, Room 406, Boston, MA 02115 (eschneider@hsph.harvard.edu).

Author Contributions: Drs Schneider, Zaslavsky, and McGlynn had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: Schneider, Nadel, Zaslavsky, and McGlynn. Acquisition of data: Schneider and McGlynn. Analysis and interpretation of data: Schneider, Nadel, Zaslavsky, and McGlynn. Drafting of the manuscript: Schneider, Zaslavsky, and McGlynn. Critical revision of the manuscript for important intellectual content: Schneider, Nadel, and McGlynn. Statistical analysis: Zaslavsky. Obtained funding: Schneider, Nadel, and McGlynn. Administrative, technical, and material support: Nadel. Study supervision: Schneider and McGlynn.

Financial Disclosure: None reported.

Funding/Support: This study was supported by grant R18 HS09473 from the Agency for Healthcare Research and Quality and the Centers for Disease Control and Prevention.

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

Additional Contributions: David Klein, PhD, provided diligent analytic assistance.

REFERENCES


