Alternative Explanations for Poor Report Card Performance

CONTEXT. Many managed care organizations grade physician groups with “report cards” developed from administrative data sets and chart reviews.

OBJECTIVE. To investigate the accuracy of five report cards on a single group practice.

DESIGN. Determination of report card accuracy by using the practice capitation list and a review of the patients’ medical records.


RESULTS. Four major problems were uncovered. First, four of the five report cards included patients who were enrolled in our practice for only a portion of the reporting year (for the four report cards, the proportion of partial-year enrollees was 8%, 15%, 23%, and 100%). Second, there was a considerable number of false-positive diagnoses in the administrative algorithms. Eight of the 61 patients labeled with hypertension did not have this condition (error rate, 14%). Other error rates were 44% for coronary artery disease, 50% for congestive heart failure, 33% for atrial fibrillation, and 0% for diabetes. Third, the administrative data often failed to capture laboratory data. Laboratory performance measures for patients with diabetes (hemoglobin A1c and cholesterol measurement and screening for microalbuminuria) were 3 to 10 times higher when assessed by chart review. Finally, the uniformly small sample sizes used in the report cards make the estimates of performance imprecise. No report card reported 95% CIs.

CONCLUSION. Five report cards on a group practice contained methodologic problems that led to systematic underestimation of the practice’s performance. Larger surveys are needed to determine the accuracy of report cards in current use.

Most managed health care plans regularly send reports to participating primary care physicians. These reports gauge physicians’ hospital utilization and specialty referral patterns. In addition, many plans evaluate physicians’ clinical performance using standard HEDIS (Health Plan Employer Data and Information Set) measures and other indicators. These “report cards” present clinical opportunities. Physicians can learn how completely (or incompletely) they vaccinate their elderly patients or how often their patients with asthma are seen in emergency departments. By identifying areas for improvement, report cards can foster better care.

Report cards may also affect a physician’s reputation. In California, two health plans released physician-specific ratings from clinical audits and patient satisfaction surveys to members. In addition, some plans publish their physician ratings on the Internet. The Federation of State Medical Boards recommends that state medical boards consider using health plan report cards to help identify substandard physicians. The reliability and accuracy of these data, however, are not well known. Poor-quality report cards may unnecessarily impugn good care.

The abstract of this paper is available at ecp.acponline.org.
I describe one practice’s experience with report cards. My colleagues and I were disturbed to learn about our own disappointing clinical performance. One report card stated that our physicians omitted blood pressure measurements for 36% of enrollees. Another stated that we administered anticoagulation therapy to only 12% of patients with atrial fibrillation. These reports prompted chart reviews. My colleagues and I wished to understand how we had failed to accomplish the intervention or, alternatively, why the report card failed to capture the care.

**Methods**

**Setting**

The practice is located in downtown Philadelphia, Pennsylvania. All 19 physicians in the practice are trained in internal medicine and are affiliated with Jefferson Medical College. At the time of the study, the practice had contracts with three capitated health plans, three preferred provider organizations, and one point-of-service plan. Patients from these plans constituted over half the practice in both volume and income and represented most of its growth.

Two of the capitated plans produced a total of five report cards between 1994 and 1997. As shown in Table 1, each of the five report cards used a 1-year study period and focused on either preventive care or disease management. The health plans used nurse auditors to gather data from outpatient records or gleaned it from administrative sources, including claims, pharmacy, and laboratory data. Results affected capitation rates; poor performers experienced a decrease in their monthly payments.

**Determination of Accuracy**

The two plans provided patient names for each report card. I reviewed both the patient’s enrollment status (to address the question, “Were we responsible for this patient at the time of the report card?”) and the outpatient chart (to evaluate the performance measure itself). The discrepancies between the report card and our records were noted and catalogued by using the following categories.

**Partial-Year Enrollment**

The practice receives a monthly capitation list from the plans that notes the date on which patients selected their primary provider. All patients who were included in the report card and were enrolled in the practice for only part of the 12-month survey period were classified as partial-year enrollees.

**Diagnostic Misclassification**

The two disease management report cards relied solely on administrative data to establish the diagnosis. The diagnosis for each patient included in the report card was reevaluated in the medical record. The entire outpatient record was reviewed to evaluate five diagnoses.

A diagnosis of hypertension was considered erroneous if the chart lacked three elevated blood pressures, the problem list and notes made no mention of past hypertension, and the patient was not receiving pharmacotherapy for hypertension. A diagnosis of coronary artery disease was considered erroneous if the patient had normal coronary arteries at cardiac catheterization or a normal noninvasive cardiac workup or if a cardiologist decided that the patient’s symptoms were non-cardiac and discontinued treatment. A diagnosis of congestive heart failure was considered erroneous if the ejection fraction exceeded 50% and there was no evidence that the physician considered congestive heart failure as an explanation for the patient’s symptoms. A diagnosis of atrial fibrillation was considered erroneous when no past or present evidence of atrial fibrillation was recorded in the chart. A diagnosis of diabetes was considered erroneous if all random blood sugar levels and hemoglobin A1c levels were normal.

**Missing Laboratory Data**

The diabetes report card used three laboratory tests as performance measures (hemoglobin A1c and cholesterol measurement and screening for microalbuminuria). A test was recorded as being present when a laboratory result obtained during the 12 months of the study period was in the chart. These data were compared with the report card results.
95% CIs were calculated on the proportions in one report card measuring preventive services. The following equation was used:

\[
95\% \text{ CI} = p \pm 1.96 \sqrt{pq/n}
\]

where \(n\) is the total number of patients in the sample, \(p\) is the proportion of the sample meeting the performance measure, and \(q = 1 - p\).

**Results**

**Partial-Year Enrollment**

Four of the five report cards included partial-year enrollees in the patient sample. Partial-year enrollees are a particular concern in the performance measurement of annually recommended services (e.g., flu shots, fecal occult blood screening, and ophthalmologic evaluation in diabetic patients). It is misleading to hold a practice responsible for an annual flu shot, for example, in a patient who has been enrolled for only 2 months. As shown in Table 1, the proportion of sampled patients who were not enrolled for the full 12 months of the survey period varied from 0% to 100%.

**Diagnostic Misclassification**

The chart could not confirm many of the administrative diagnoses for patients in the cardiac care report card. As shown in Figure 1, the accuracy of the administrative diagnoses varied depending on the specific diagnosis. The rate of erroneous diagnoses ranged from 50% (5 of 10 patients) in congestive heart failure to 0% (0 of 52 patients) in diabetes (although 3 of the 52 patients developed diabetes during the survey year). Table 2 shows the results of the chart audit for patients in whom the administrative algorithm misdiagnosed coronary artery disease, hypertension, and atrial fibrillation.

**Missing Data**

Some report cards rely on laboratory data obtained from administrative sources. As shown in Table 3, the proportion of diabetic patients who met laboratory performance measures was 3- to 10-fold higher on chart audit than in the administrative database. When partial-year enrollees were eliminated, the practice performance on diabetic indices improved further.

![Figure 1. Proportion of report card diagnoses that were not confirmed by chart review.](image)
Precision

The report cards were based on sample sizes ranging from 22 to 89 patients. These samples were often further subdivided by age (preventive samples) or disease (cardiac sample). Thus, the CIs surrounding the proportions were often wide, limiting the reliability of any single quality measurement. Table 4 provides one example of this problem, using routine preventive performance measures for enrollees 40 to 64 years of age. None of the report cards reported 95% CIs for any measurement.

Discussion

The National Committee for Quality Assurance judges health plan performance on the basis of data from members who have been continuously enrolled in the plan for 12 months.\(^5\) However, in this study, four of five report cards judged a practice’s performance without such exclusions. Partial-year enrollees confound report card results by measuring annual indices in patients whom a practice has treated for less than 1 year.
Our own experience is illustrative. In 1994, a report card placed our practice in the lowest 8% of physicians for quality. The plan reported that the practice neglected all clinical indicators in over one third of sampled patients; as a result, our capitation rate plummeted. The chart audit, however, revealed most of these patients had never visited or called the office, probably because they had been enrolled in the practice for only a few months. None had been enrolled for a full year. After the chart audit, the plan adjusted the quality rating; they agreed that the partial-year enrollees skewed the report card results. Because plan members change physicians, health plans should include only enrollees who have been capitated to the practice for a full year when measuring a practice’s performance.

This study also shows how other methodologic problems can erroneously indicate poor performance. The cardiac report card stated that the practice prescribed β-blockers in only 33% of patients with coronary artery disease. However, the chart audit found that the diagnosis of coronary artery disease was erroneous in 44% of patients, perhaps because ICD-9 (International Classification of Diseases, ninth revision) codes lack a designation for the “rule-out” diagnosis. Eliminating the 11 patients with a misdiagnosed disease leaves 14 patients with coronary artery disease and a β-blocker prescription rate of 50%. The algorithm’s diagnostic errors obscure the practice’s real behavior and created a sample that was too small for meaningful comparisons.

The accuracy of administrative algorithms for determining diagnoses may depend on the specific diagnosis. In hospitalized patients, claims data missed 52% of patients with congestive heart failure and 32% of

<table>
<thead>
<tr>
<th>TABLE 3</th>
<th>The Effect of Measurement Method on Diabetes Care Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIABETES PERFORMANCE MEASURE</td>
<td>REPORT CARD RESULTS (n = 52)</td>
</tr>
<tr>
<td></td>
<td>ALL PATIENTS (n = 52)</td>
</tr>
<tr>
<td>Hemoglobin A₁c measurement (biannually)</td>
<td>5.8%</td>
</tr>
<tr>
<td>Cholesterol measurement (annually)</td>
<td>11.6%</td>
</tr>
<tr>
<td>Screening for microalbuminuria (annually)</td>
<td>7.8%</td>
</tr>
</tbody>
</table>

*Twelve partial-year enrollees were excluded from the analysis.
patients with hypertension, although clinical and claims data correlated strongly for the diagnosis of diabetes. In addition, Medicare claims data and chart audits concur on the proportion of diabetic patients in whom hemoglobin A1c levels have been measured. However, in this study, the chart audit diverged dramatically from administrative data in identifying laboratory information on diabetic patients. This is explained by the fact that most of our patients are referred to a laboratory that is not connected to the health plan’s database.

This case report, although clearly limited in scope, has a simple message for physicians: Carefully scrutinize your report cards. These reports increasingly influence physician compensation, and public dissemination of report cards may unfairly tarnish the reputations of physicians. This study also provides evidence that plans need to field-test report cards to determine their accuracy. Administrative algorithms and computerized databases need to be validated with large-scale chart audits. Sample sizes should be increased to measure physician behavior more precisely. At a minimum, all report cards should report CIs.

The larger untested question is whether report cards will improve care. Although my practice achieved better scores than the plan mean, the chart reviews also confirmed genuine deficiencies. We are making efforts to improve care of patients with diabetes and are developing plans to check all patients with congestive heart failure for use of angiotensin-converting enzyme inhibitors, β-blockers, and spironolactone. Report cards, despite their inaccuracies, can provide guidance for clinical improvement.

Even if report cards were perfectly accurate, however, it remains unclear what portion of ideal performance is under physician control. Business leaders assume that physician compensation, and public dissemination of report cards may unfairly tarnish the reputations of physicians. This study also provides evidence that plans need to field-test report cards to determine their accuracy. Administrative algorithms and computerized databases need to be validated with large-scale chart audits. Sample sizes should be increased to measure physician behavior more precisely.

Take-Home Points

- Chart audits were performed to confirm the results, and the enrollment status of sampled patients was checked against capitation lists.
- Four of the five report cards included patients who were enrolled in the practice for only a portion of the reporting year.
- Other problems included diagnostic errors, missing laboratory data, and inadequate sample size.
- Although report cards may help improve medical practice, their accuracy is open to question.

### TABLE 4
59% CIs for Three Variables in a Report Card*

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>FINDINGS (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood pressure recorded</td>
<td>64% (48%–80%)</td>
</tr>
<tr>
<td>Screened for alcohol and substance abuse</td>
<td>42% (26%–58%)</td>
</tr>
<tr>
<td>Cholesterol measured</td>
<td>61% (45%–77%)</td>
</tr>
</tbody>
</table>

*Measured in enrollees 40 to 64 years of age (n = 36).

### References

### Acknowledgments
The author thanks Kenneth Epstein, MD, Christine Laine, MD, and Barbara Turner, MD, for comments on earlier versions of this manuscript.

### Correspondence
Rachel Sorokin, MD, 12 South Ninth Street, Suite 502, Jefferson Medical College, Philadelphia, PA 19107; telephone: 215-955-0733; fax: 215-923-9239; e-mail: rachel.sorokin@mail.tju.edu.