Assessing the Validity of National Quality Measures for Coronary Artery Disease Using an Electronic Health Record

Stephen D. Persell, MD, MPH; Jennifer M. Wright, MD; Jason A. Thompson, BA; Karen S. Kmetik, PhD; David W. Baker, MD, MPH

Background: Nationally endorsed, clinical performance measures are available that allow for quality reporting using electronic health records (EHRs). To our knowledge, how well they reflect actual quality of care has not been studied. We sought to evaluate the validity of performance measures for coronary artery disease (CAD) using an ambulatory EHR.

Methods: We performed a retrospective electronic medical chart review comparing automated measurement with a 2-step process of automated measurement supplemented by review of free-text notes for apparent quality failures for all patients with CAD from a large internal medicine practice using a commercial EHR. The 7 performance measures included the following: antiplatelet drug, lipid-lowering drug, β-blocker following myocardial infarction, blood pressure measurement, lipid measurement, low-density lipoprotein cholesterol control, and angiotensin-converting enzyme inhibitor or angiotensin receptor blocker for patients with diabetes mellitus or left ventricular systolic dysfunction.

Results: Performance varied from 81.6% for lipid measurement to 97.6% for blood pressure measurement based on automated measurement. A review of free-text notes for cases failing an automated measure revealed that misclassification was common and that 15% to 81% of apparent quality failures either satisfied the performance measure or met valid exclusion criteria. After including free-text data, the adherence rate ranged from 87.5% for lipid measurement and low-density lipoprotein cholesterol control to 99.2% for blood pressure measurement.

Conclusions: Profiling the quality of outpatient CAD care using data from an EHR has significant limitations. Changes in how data are routinely recorded in an EHR are needed to improve the accuracy of this type of quality measurement. Validity testing in different settings is required.

Arch Intern Med. 2006;166:2272-2277
to demonstrate the utility of these measure specifications by having several sites report standardized electronic data to a central source. However, how well electronic data will mirror patients’ true clinical care has not been extensively evaluated.

The accuracy of EHR-based measures depends on how thoroughly and accurately the data used in the measures are recorded in standardized portions of an EHR. The way clinicians use EHRs for documentation may not be ideal for capturing quality data, especially if patients receive care from multiple sources. Furthermore, many measures include explanations that allow for patient and physician exceptions for not meeting a measure, but EHRs may not contain standardized ways to document these exceptions. As quality of care improves, exceptions to the rule and miss measurement may compose a large portion of apparent quality failures.

We aimed to evaluate the validity of nationally endorsed quality measures for CAD in a large internal medicine practice using a commercial EHR and to assess the accuracy of apparent quality failures. We also sought to learn how this process could be improved.

**METHODS**

**QUALITY MEASURES**

We examined 7 performance measures for CAD included in the DOQ-IT project (**Table 1**). Six measures reflected processes of care: antiplatelet drug prescribed, lipid-lowering drug prescribed, β-blocker prescribed for patients with prior myocardial infarction, blood pressure measurement, lipid measurement, and angiotensin-converting enzyme (ACE) inhibitor or angiotensin receptor blocker (ARB) prescribed for patients with both CAD and diabetes mellitus and/or left ventricular systolic dysfunction. One measure, low-density lipoprotein cholesterol (LDL-C) level lower than 130 mg/dL (3.4 mmol/L), was an intermediate outcome.

**SETTING AND PATIENTS**

Quality measures for CAD were applied to patients from the General Internal Medicine Clinic of the Northwestern Memorial Faculty Foundation, Chicago, Ill. This clinic employs 40 full- or part-time internal medicine physicians and provides more than 41 000 patient visits annually. The practice uses an EHR (Epic; Epic Systems Corporation, Madison, Wis) for all clinical documentation including note writing, prescribing, and reporting test results. Diagnosis codes based on the International Classification of Diseases, Ninth Revision (ICD-9), are linked to test ordering and office visits and may be entered into patients’ medical history or problem lists.

Patients were selected for study inclusion based on criteria from DOQ-IT CAD narrative version 1.2, May 2005 (CMS, Baltimore, Md) (available from the authors by request). Patients with visit, problem list, or medical history diagnosis of CAD (ICD-9 codes 414.00-414.07, 414.8, 414.9, 410.00-410.92, 412, V45.81, V45.82, 411.0-411.89, or 413.0-413.9) were included. Patients had to be at least 18 years of age and have 2 or more office visits within the study interval (calendar year 2004). Coronary artery disease did not have to be the visit diagnosis.

**APPLICATION OF ELECTRONIC QUALITY MEASURES**

We determined whether patients with CAD met each performance measure. As specified, patients who did not meet a quality measure were excluded from the denominator for that measure if electronic exclusion criteria were met. For the ACE inhibitor/ARB measure, we used diagnosis codes for diabetes mellitus and a preestablished registry to classify patients as having left ventricular systolic dysfunction if the left ventricular ejection fraction was estimated to be lower than 40% by echocardiography, angiography, or nuclear scintigraphy.

**MEDICAL CHART ABSTRACTION**

Two physicians (S.D.P. or J.M.W.) performed structured abstractions of electronic medical charts for patients who did not meet 1 or more measure. This record review included the coded and free-text portions of the EHR. We sought to determine if 2 types of misclassification error were present: (1) failure to detect that a patient met quality criteria and (2) failure to detect exclusion criteria. For patients who did not meet the blood pressure measure, a nonphysician reviewed the last office visit during the study interval to determine if blood pressure was recorded in the text of the office note.

---

**Table 1. Summary of Coronary Artery Disease (CAD) Quality Measures**

<table>
<thead>
<tr>
<th>Quality Measure</th>
<th>Inclusion</th>
<th>Valid Exclusions When Measure Not Met</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antiplatelet drug (aspirin or clopidogrel)</td>
<td>All patients with CAD</td>
<td>Medical contraindication, allergy, adverse effect, use of dipyridamole or ticlopidine, patient or medical reason to forgo</td>
</tr>
<tr>
<td>Lipid-lowering drug</td>
<td>All patients with CAD</td>
<td>Medical contraindication, allergy, adverse effect, pregnancy, LDL-C level less than 130 mg/dL (&lt;3.4 mmol/L), patient or medical reason to forgo</td>
</tr>
<tr>
<td>β-Blocker</td>
<td>Patients with both CAD and prior myocardial infarction</td>
<td>Medical contraindication, allergy, adverse effect, patient or medical reason to forgo</td>
</tr>
<tr>
<td>Blood pressure measurement†</td>
<td>All patients with CAD</td>
<td>Patient reason to forgo blood pressure measurement</td>
</tr>
<tr>
<td>Lipid measurement</td>
<td>All patients with CAD</td>
<td>Patient or medical reason to forgo lipid measurement</td>
</tr>
<tr>
<td>LDL-C &lt;130 mg/dL (≤3.4 mmol/L)</td>
<td>All patients with both CAD and LDL-C measurement</td>
<td>None</td>
</tr>
<tr>
<td>ACE inhibitor/ARB</td>
<td>All patients with both CAD and diabetes mellitus or left ventricular systolic dysfunction</td>
<td>Medical contraindication, allergy, adverse effect, pregnancy, patient or physician reason to forgo</td>
</tr>
</tbody>
</table>

Abbreviations: ACE, angiotensin-converting enzyme; ARB, angiotensin receptor blocker; LDL-C, low-density lipoprotein cholesterol.

*Measures are from the Doctor’s Office Quality–Information Technology narrative version 1.2, May 2005 (Centers for Medicare and Medicaid Services, Baltimore, Md) (available from the authors by request).

†At the last office visit during the study interval with a physician, nurse practitioner, or physician assistant.
When free-text portions of the EHR demonstrated that quality criteria were met, we classified patients as meeting that measure on chart review. When patients who did not meet a measure had a reason for exclusion noted in the free-text portion of the EHR or did not have confirmed CAD, that patient was classified as excluded on chart review. We confirmed the CAD diagnosis if a patient had a history of a clinical myocardial infarction, prior coronary revascularization, or chronic stable angina. Patients who met none of these criteria were considered to have confirmed CAD if coronary angiography revealed any local luminal stenosis or when angiography was not performed, by an exercise or pharmacologic stress test, or if high-resolution computed tomography of the coronary arteries showed at least moderate abnormalities suggestive of CAD. Since not all test results were available in the EHR, a physicians' description of these findings was acceptable to confirm the diagnosis. When a single reviewer was uncertain of the diagnosis of CAD, both reviewers discussed the case and reached consensus.

For measures that allowed for exclusion when a medical or patient reason for forgoing a treatment was present, we considered any of the following descriptions to meet exclusion criteria: a clinical rationale for forgoing treatment or treatment intolerance, nonadherence, refusal, or unaffordability. For the antiplatelet drug measure, we calculated the success rate after excluding patients who were receiving oral anticoagulation and failed to meet the measure because the risks and benefits of adding aspirin in this setting differ considerably from using aspirin alone. For the lipid measurement measure, because it could reflect a patient's reason to forgo testing, we excluded patients for whom a physician ordered the test but it was not performed. For the blood pressure measure, we excluded patients who failed to meet this measure if blood pressure measurement was attempted but could not be obtained.

Of the medical charts, 15% were abstracted by both physician reviewers. Interrater reliability for determining whether a patient was misclassified based on physician review using the κ statistic was good and ranged from 0.62 to 0.85 for individual measures and was 0.78 for all measures.

### ADDITIONAL MEASURE CRITERIA

For the LDL-C control measure, we thought exclusions that were not part of the measure criteria could potentially explain why some patients did not meet the measure. We believed that patients unable or unwilling to use cholesterol-lowering medication would be unlikely to reach the LDL-C control measure. Therefore, we calculated the rate of LDL-C control after excluding the patients meeting the exclusion criteria for the lipid-lowering drug measure.

### ANALYTIC APPROACH

We calculated performance rates of the quality measures as follows:

**Number Meeting Criteria/**

(Number Meeting Criteria + Number Not Meeting Criteria With No Exclusion Criteria).

We repeated these calculations after reclassifying patients based on chart review. Patients judged not to have CAD were removed from all denominators for the calculation of the rates after chart review. We calculated the difference between measures calculated using the 2-step process and the entirely automated process.

This study was approved by the institutional review board of Northwestern University (Chicago, Ill).

### RESULTS

Our search identified 1006 patients with CAD diagnosis codes and 2 or more office visits in 2004. The mean (SD) age was 65.2 (11.4) years; 43.4% were female; 27.6% were black, 49.2% were white, 5.1% were Latino, and 18.1% were another race or of unknown race. Diagnosis codes were present for myocardial infarction for 19.0%, prior coronary artery bypass graft surgery for 13.0%, and percutaneous coronary revascularization for 12.4%. Hypertension and diabetes were common, present in 46.7% and 22.1%, respectively.

### AUTOMATED MEASUREMENT COMPARED WITH MANUAL REVIEW OF ELECTRONIC MEDICAL CHARTS

The number of patients eligible for each measure varied from 134 for β-blocker prescribed after myocardial infarction to 1006 for blood pressure and lipid measurement (Table 2). By automated measurement, 454 patients (45.1%) did not meet 1 or more quality measure for which they were eligible. The percentage of patients who satisfied the individual measures ranged from 81.6% for lipid measurement to 97.6% for blood pressure measurement (Table 2). Physician chart review of patients who did not meet a quality measure detected additional cases in which the numerator or exclusion criteria went undetected (Table 3). Therefore the number meeting performance criteria increased and the number who remained eligible for each measure decreased. Of 454 patients who failed 1 or more measure, 61 (13.4%) did not have confirmed CAD (as described in the “Methods” section). Recalculation of the quality measures using the 2-step process yielded higher success rates than the automated results (Table 4). After chart review, success rates were 1.5 to 14.3 percentage points higher and varied from 87.1% for LDL-C control (<130 mg/dL [<3.4 mmol/L]) to 99.1% for blood pressure measurement at the last office visit (Table 4).

### MEASURE WITH ADDITIONAL CRITERIA ADDED

We calculated the LDL-C control rate after excluding patients who meet exclusion criteria for the lipid-lowering drug measure.

---

### Table 2. Results of Automated Performance Assessment for Coronary Artery Disease

<table>
<thead>
<tr>
<th>Quality Measure</th>
<th>Criteria Met, No.</th>
<th>Eligible, No.</th>
<th>Performance, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antiplatelet drug</td>
<td>544</td>
<td>664</td>
<td>81.9</td>
</tr>
<tr>
<td>Lipid-lowering drug</td>
<td>830</td>
<td>897</td>
<td>92.5</td>
</tr>
<tr>
<td>β-Blocker after MI</td>
<td>111</td>
<td>134</td>
<td>82.8</td>
</tr>
<tr>
<td>Blood pressure measured</td>
<td>982</td>
<td>1006</td>
<td>97.6</td>
</tr>
<tr>
<td>Lipid measurement</td>
<td>821</td>
<td>1006</td>
<td>81.6</td>
</tr>
<tr>
<td>LDL-C control</td>
<td>700</td>
<td>821</td>
<td>85.3</td>
</tr>
<tr>
<td>ACE inhibitor/ARB</td>
<td>241</td>
<td>283</td>
<td>85.2</td>
</tr>
</tbody>
</table>

Abbreviations: ACE, angiotensin-converting enzyme; ARB, angiotensin receptor blocker; LDL-C, low-density lipoprotein cholesterol; MI, myocardial infarction.
valid exclusions were also common. Electronic searches. Erroneous CAD diagnoses and missed detected patients meeting the measure who were missed by parent failures for each measure. In all cases we de-
misclassification with these patients excluded was 49%.

exclusion criteria (exclusions for lipid-lowering drug measure). The rate of (electronic review met the criteria of having an LDL-C level under 130 mg/dL measure because they did not have an LDL-C measurement detected on infarction.

receptor blocker; LDL-C, low-density lipoprotein cholesterol; MI, myocardial infarction.

rather than true quality failures. Implementing measures ity failures were frequently due to misclassification errors of outpatient CAD care at a site where the rates of satisfying performance criteria exceeded 80%, apparent quality failures were frequently due to misclassification errors rather than true quality failures. Implementing measures that result in this form of misclassification could increase physician opposition to performance measurement for provider accountability or payment. Fear of being labeled as delivering poor care may discourage physicians from caring for medically challenging, economically disadvantaged, or nonadherent patients and could encourage clinicians to disregard patients’ preferences or ignore their comorbid conditions when making treatment recommendations. Furthermore, if care improves overall without overcoming measurement problems, the majority of apparent quality failures will actually be “false positives”—patients who appear to have not received indicated care who actually had a contraindication, exclusion, or unwilling-ness to follow a recommendation.

Our study testing CAD quality measures using outpatient data helps identify the reasons these measures misclassify patients as false quality failures. Addressing 4 general types of error could substantially remedy the measurement problems we encountered. First, as in prior studies, diagnoses were often used incorrectly. As a result, patients were improperly labeled as having CAD. Patients who had chest symptoms at one point in time or had a test ordered to exclude CAD may have been given the diagnosis of “angina pectoris” or “coronary atherosclerosis” even though subsequent evaluation did not suggest CAD. Without a method to overrule prior erroneous diagnoses, these patients remain on a list of patients eligible for CAD quality measures. Creation of new “diagnostic codes” (i.e., “er-

evasive CAD diagnosis prior to this date”) that could be used to override codes used incorrectly in the past is a potential solution to this problem.

Second, data that would have fulfilled quality criteria were not always documented in searchable portions of patient records. For example, physicians often wrote that they told patients to use aspirin but did not enter it into the standardized medication list because it does not require a prescription. Also, some prescription medications clearly mentioned in office notes were missing from
drug measure. Forty-one patients failing to meet this measure met the exclusion criteria for the lipid-lowering drug measure. When recalculated using these additional criteria, the performance rate was 91.8%, a 6.5 percentage point increase.

REASONS FOR MISCLASSIFICATION

Table 5 gives the reasons for misclassification of apparent failures for each measure. In all cases we detected patients meeting the measure who were missed by electronic searches. Erroneous CAD diagnoses and missed valid exclusions were also common.

COMMENT

Quality measures that use electronic clinical information systems may enable the assessment of clinical performance in ways that have not been possible using paper-based records or electronic data generated for nonclinical purposes. Since they are used for clinical documentation and not merely billing, EHRs may allow for more accurate quality measurement than can be accomplished using administrative data alone, while keeping data collection efficient. Despite these advantages, our study shows that critical information required to measure quality may still not be captured in a way that can be used by algorithms designed to measure quality of care using EHRs. As a result, the quality measures currently undergoing demonstration by CMS could yield inaccurate estimates of quality.

When we used EHR-based measures to assess the quality of outpatient CAD care at a site where the rates of satisfying performance criteria exceeded 80%, apparent quality failures were frequently due to misclassification errors rather than true quality failures. Implementing measures
the medication list. Patients from the study practice receive care from many other health care providers. We found cases in which physicians recorded data about prescriptions or laboratory results from physicians outside the practice that would have met the measure criteria. Because this information was located in the free-text portion of office notes, automated searches did not detect it. Configuring EHRs to allow for easy standardized documentation of data collected from other sources could ameliorate this problem.

Third, allowable exceptions for quality measure failures were not reliably captured by automated searches of the EHR. Even though the measures include exclusion criteria concerning patient or medical reasons for forgoing treatment, without a simple, standardized way to capture these exclusions, they cannot be used as part of automated measures. One potential solution would be creating new standard codes that explain exceptions or patient preferences for important chronic disease treatment decisions (eg, “hypotension due to ACE inhibitor” or “recommended β-blocker and patient declined”). These codes could be placed on patients’ problem lists or associated with clinical encounters. Alternatively, electronic forms within EHRs could be used that explicitly capture the information used by physicians to make chronic disease management decisions in standardized fields so that it could be readily used for quality measurement or improvement.

Lastly, the LDL-C measure did not allow for exceptions that explained a large number of apparent failures. We determined the success rate for the LDL-C control measure after allowing for additional exclusions when there was a patient factor that prevented the measure from being achieved that was not within the physician’s direct control. Because controlling the LDL-C of patients who are unable or unwilling to use cholesterol-lowering medication may not be possible, we chose to calculate the performance of LDL-C control allowing for the same exceptions from the lipid-lowering drug use measure. While this decision is controversial, allowing these exceptions may be necessary to increase physician acceptance of this kind of measurement and to avoid the creation of perverse incentives to steer physicians away from nonadherent or medically complex patients. In our study population, once we allowed for these exceptions, the performance rate for this measure was over 90%. Continued refinement of measure specifications will likely be needed as medical evidence evolves.

### Table 5. Types of Misclassification for Patients Failing to Meet Quality Measures

<table>
<thead>
<tr>
<th>Quality Measure</th>
<th>Patients (No.) Meeting Measure Criteria on Medical Chart Review</th>
<th>Patients (No.) Meeting Exclusion Criteria on Medical Chart Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antiplatelet drug</td>
<td>Prescribed aspirin (38)</td>
<td>No CAD (31), taking warfarin (11), contraindication (10), adverse reaction (2), terminally ill (2), refused (2), nonadherent (1)</td>
</tr>
<tr>
<td>Lipid-lowering drug</td>
<td>Prescribed lipid-lowering drug (5)</td>
<td>No CAD (29), adverse reaction (5), contraindication (1), cannot afford (1), terminally ill (1), refused (1)</td>
</tr>
<tr>
<td>β-Blocker after MI</td>
<td>Prescribed β-blocker (1)</td>
<td>No MI (6), contraindication (1), adverse reaction (1), refused (1), nonadherent (1)</td>
</tr>
<tr>
<td>Blood pressure measured</td>
<td>Blood pressure recorded (11)</td>
<td>No CAD (5), attempted but unable to obtain (2), refused blood pressure measurement (1)</td>
</tr>
<tr>
<td>Lipid measurement</td>
<td>Lipid measurement recorded (16)</td>
<td>No CAD (25), test ordered, not done (30)</td>
</tr>
<tr>
<td>LDL-C control</td>
<td>LDL-C level below goal (15)</td>
<td>No CAD (18), contraindication* (3), adverse reaction* (20), cannot afford* (4), refused* (7), nonadherent* (7)</td>
</tr>
<tr>
<td>ACE inhibitor/ARB</td>
<td>Prescribed ACE inhibitor or angiotensin-converting enzyme inhibitor (4)</td>
<td>No CAD (2), adverse reaction (4), terminally ill (1), refused (3)</td>
</tr>
</tbody>
</table>

Abbreviations: ACE, angiotensin-converting enzyme; ARB, angiotensin receptor blocker; CAD, coronary artery disease; LDL-C, low-density lipoprotein cholesterol; MI, myocardial infarction.
*Additional exclusion criteria.
Our study has several limitations. The fact that we assessed the CAD measures for only 1 group practice with a single EHR limits the generalizability of our findings. We suspect that there could be considerable variation in the accuracy of these measures across practices, and we cannot determine from a single practice the extent to which these measurement errors could influence numerical comparisons between practices. Testing the validity of automated measures in multiple practice settings would provide a better understanding of how measurement error would affect quality comparisons between practices or health plans. Another potential source of error in our methodology is that we chose to allow inclusion of patients with CAD diagnosis codes who only had evidence of subclinical CAD. Therefore, some patients remained eligible for CAD quality measures on chart review who had not yet developed clinically evident CAD. If we had excluded these patients, we would have excluded more patients failing quality measures and found an even wider difference between the two sets of measurements. We did not determine if some patients meeting all quality measures did not in fact have CAD, though it is unlikely that this would have caused more than minimal changes to our findings. Lastly, we were unable to determine if there were patients in this population with clinically diagnosed CAD who did not have any CAD diagnosis code used.

Measuring outpatient quality of care is a complicated endeavor. Our findings suggest that if high-performing sites were compared using quality measures derived from EHRs, apparent numerical differences could be largely due to measurement problems rather than quality differences. Formally assessing how well automated measures perform prior to their adoption for pay-for-performance and public accountability seems to be a worthy exercise. Errors in diagnostic coding, documentation of care, the capture of exclusions, and accounting for valid reasons for quality failures present important obstacles to accurate measurement. Changes in the way clinicians document chronic disease care, the addition of standardized codes to explain management decisions, and designing EHRs to better serve chronic disease care may make EHR-based CAD quality measures more accurate. At present, it would be premature to use these electronic measures for public comparison or reimbursement.

Accepted for Publication: August 1, 2006.

Correspondence: Stephen D. Persell, MD, MPH, Division of General Internal Medicine, Northwestern University, 676 N St Clair St, Suite 200, Chicago, IL 60611-2927 (spersell@nmff.org).

Author Contributions: Dr Persell had full access to all the data and takes responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: Persell and Baker. Acquisition of data: Persell, Wright, and Thompson. Analysis and interpretation of data: Persell, Kmetik, and Baker. Drafting of the manuscript: Persell. Critical revision of the manuscript for important intellectual content: Wright, Thompson, Kmetik, and Baker. Statistical analysis: Persell, Thompson. Obtained funding: Kmetik. Administrative, technical, and material support: Thompson and Kmetik. Study supervision: Baker. Drs Persell and Baker designed the study and conducted the data analysis independently from the American Medical Association. Dr Kmetik provided input into the final manuscript.

Financial Disclosure: None reported.

Funding/Sponsor: Funding for this study was provided by grant 5 U18HS13690-02 from the Agency for Healthcare Research and Quality.

Role of the Sponsor: The funding agency was not involved in the design or the decision to submit for publication.

Acknowledgment: We thank Heidi Bossley, MSN, MBA, for her valuable contribution to this project.

REFERENCES


