

Automated Review of Electronic Health Records to Assess Quality of Care for Outpatients with Heart Failure

David W. Baker, MD, MPH; Stephen D. Persell, MD, MPH; Jason A. Thompson, BA; Neilesh S. Soman, MD, MBA; Karen M. Burgner, MD; David Liss, BA; and Karen S. Kmetik, PhD

Background: Electronic health records (EHRs) may be used to assess quality of care.

Objective: To evaluate the accuracy of automated review of EHR data to measure quality of care for outpatients with heart failure.

Design: Observational study of quality of care for heart failure comparing automated review of EHR data with automated review followed by manual review of electronic notes for patients with apparent quality deficits (hybrid review).

Setting: An academic general internal medicine clinic with several years' experience using a commercial EHR.

Patients: 517 adults with a qualifying International Classification of Diseases, Ninth Revision, diagnosis of heart failure in their EHR data and 2 or more clinic visits over the past 18 months.

Measurements: Left ventricular ejection fraction (LVEF), prescription of a β -blocker and an angiotensin-converting enzyme (ACE) inhibitor or angiotensin-receptor blocker (ARB) for patients with left ventricular systolic dysfunction (LVEF <0.40) and prescription of warfarin for patients with comorbid atrial fibrillation.

Results: Performance based on automated review of EHR data was similar to that based on hybrid review for assessing LVEF measure-

ment (94.6% vs. 97.3%), prescription of β -blockers (90.9% vs. 92.8%), and prescription of ACE inhibitors or ARBs (93.9% vs. 98.7%). However, performance based on automated review was lower than that based on hybrid review for prescription of warfarin for atrial fibrillation (70.4% vs. 93.6%), primarily because automated review did not detect documentation of accepted reasons for not prescribing warfarin.

Limitations: The findings may not be applicable to other practices and other EHRs. The authors used EHR data to identify eligible patients, so the study may have excluded some patients with heart failure. Patient charts were manually reviewed only if a provider appeared to fail a quality measure on automated review and did not determine the sensitivity and specificity of automated review according to standard definitions.

Conclusions: Automated review of EHR data to measure the quality of care of outpatients with heart failure missed many exclusion criteria for medications documented only in providers' notes. As a result, it sometimes underestimated performance on medication-based quality measures.

Ann Intern Med. 2007;146:270-277.

For author affiliations, see end of text.

www.annals.org

The Institute of Medicine and the Medicare Payment Advisory Commission have called for government payers to publicly report providers' clinical quality data and increase payments for health care providers who deliver higher-quality care (that is, pay-for-performance) (1–4). The Centers for Medicare & Medicaid Services (CMS) has already taken steps toward these goals with the Nursing Home Quality Initiative, the Home Health Quality Initiative, and the Hospital Quality Initiative (5, 6). More recently, the CMS announced the Physician Focused Quality Initiative, which will assess the quality of outpatient medical care for key illnesses and clinical conditions and “investigate the concept of payment for performance” (7). Assessing the quality of outpatient care, especially primary care, has unique challenges. Unlike quality-of-care assessment for hospitalized patients, information relevant to outpatient quality measures may be documented in past encounters and may not easily be identifiable, especially for patients with multiple comorbid conditions and many past visits (8, 9). In addition, patients often see several different physicians, and it may not be feasible to collect information for quality measurement from other providers.

Electronic health record (EHR) systems have the potential to overcome some of these obstacles in measuring quality of care in office practices (9). The National Health Service in the United Kingdom has begun to use data automatically extracted from EHRs in family practices

throughout the country to measure providers' performance on 146 quality indicators relating to clinical care for 10 chronic diseases, organization of care, and patient experience; practices are then paid according to the proportion of patients for whom they achieve each target (10). However, few studies have examined the accuracy of automated review of data from EHR systems to measure the quality of outpatient care (11–13). Studies from the U.S. Department of Veterans Affairs have found that the success rates for checking a patient's hemoglobin A_{1c} level, low-density lipoprotein cholesterol level, and blood pressure are higher on review of clinical notes compared with automated data from an EHR (12, 13). We have previously demonstrated a high frequency of misclassification of quality measures for coronary artery disease when automated review is compared with automated review in addition to manual review of providers' free-text notes (14).

See also:

Print

Editors' Notes 271

Web-Only

Appendix

Conversion of tables into slides

The American Medical Association (AMA)–convened Physician Consortium for Performance Improvement (PCPI), in collaboration with the American College of Cardiology (ACC) and the American Heart Association (AHA), have developed physician performance measures, including measures of medication use for assessing the quality of care for patients with heart failure (15, 16). The CMS, together with PCPI, developed specifications to apply these quality measures using EHR systems (17). The CMS Doctor’s Office Quality-Information Technology (DOQ-IT) project aims to demonstrate the utility of these measure specifications (17). We undertook this study to evaluate the accuracy of an automated system that used data from an EHR system to measure quality according to the available CMS/PCPI specifications for patients with heart failure.

METHODS

Study Site and Patient Eligibility

The institutional review board at Northwestern University approved the study. We used data from the general internal medicine clinic of the Northwestern Medical Faculty Foundation. This clinic has 41 faculty members who practice full-time or part-time and 103 residents who have their own panel of clinic patients. All physicians use an EHR for all clinical encounters (EpicCare, version 7.0, Epic Systems Corporation, Madison, Wisconsin); most faculty members had 3 or more years’ experience using the EHR. History and physical examination data are usually entered as free text. The EHR has discrete fields for medical history, surgical history, allergies, social history, current and past medications (including date of order and discontinuation), vital signs, encounter diagnoses, and a problem list. Diagnosis names are linked to International Classification of Diseases, Ninth Revision (ICD-9), codes. Symptom codes (for example, “shortness of breath”) can be used if a diagnosis has not been established. There is no limit to the number of diagnoses that can be entered for a single encounter.

We included patients if they 1) had 1 or more ICD-9 codes for heart failure (**Appendix**, available at www.annals.org) listed in the problem list, medical history, or encounter diagnoses; 2) were seen 2 or more times in the general internal medicine clinic from 1 July 2003 to 31 December 2004; and 3) were 18 years of age or older. We did not attempt to identify patients who had heart failure but did not have this diagnosis entered into the EHR.

Validation of Heart Failure Diagnosis

Two investigators reviewed the physician notes of all patients who had a diagnosis of heart failure on their problem list or medical history but no encounter diagnoses ($n = 28$) to determine whether the treating physician thought the patient had previous heart failure and why it was not listed as an encounter diagnosis. We also reviewed charts for patients who had only a single-encounter diag-

Context

Searching electronic health records (EHRs) may offer a convenient way to measure quality of care.

Contribution

The authors compared automated review of EHR fields with automated review followed by manual review of the EHR to assess the quality of care of outpatients with heart failure. They found that automated review sometimes underestimated how often warfarin was prescribed to patients with atrial fibrillation because it missed documentation of legitimate reasons for not prescribing warfarin.

Cautions

The researchers did not measure how many quality measures identified by automated queries were false positives.

Implications

Automated review of the EHR missed many exclusion criteria for warfarin documented only in providers’ notes. As a result, it underestimated provider performance.

—The Editors

nosis of heart failure ($n = 66$) to determine how often heart failure was used as an encounter diagnosis to justify a diagnostic test for a patient with symptoms that could be due to heart failure.

Quality-of-Care Measures

We selected 4 of the 11 PCPI measures for this study: measurement of left ventricular ejection fraction (LVEF) for all patients with diagnosed heart failure, β -blockers for patients whose measured LVEF was less than 0.40, angiotensin-converting enzyme (ACE) inhibitors or angiotensin receptor-blockers (ARBs) for patients whose measured LVEF was less than 0.40, and warfarin for patients with heart failure and comorbid atrial fibrillation. The remaining PCPI measures were not examined because of very high rates of adherence (blood pressure and weight monitoring), because of lack of discrete electronic fields to query (assessment of clinical symptoms and signs of volume overload, assessment of activity level, and assessment of patient education), or because the measures only applied to incident cases of heart failure (laboratory testing).

Measurement of Left Ventricular Ejection Fraction

Because LVEF is not stored as a separate data field that can be queried automatically, we queried the EHR to identify all diagnostic tests that routinely measure LVEF (echocardiography, nuclear scintigraphy, and cardiac catheterization with ventriculography) and manually reviewed the test results and recorded the values when present. If no test results were available, the electronic physician notes were reviewed by 1 of the investigators to determine whether a treating physician had recorded an LVEF in the initial

Table 1. Patient Characteristics*

| Characteristic | Value |
|--|-------------|
| Location of Heart Failure Diagnosis Codes, <i>n</i> (%) | |
| Encounter diagnosis | 489 (94.6) |
| Encounter diagnosis only | 115 (22.2) |
| Problem list | 367 (71.0) |
| Problem list only | 4 (0.8) |
| Medical history | 191 (36.9) |
| Medical history only | 18 (3.5) |
| Mean visits (SD) to GIM clinic in past 18 months, <i>n</i> | 6.4 (4.6) |
| Mean age (SD), <i>y</i> | 65.1 (14.5) |
| Female, <i>n</i> (%) | 268 (51.8) |
| Race/ethnicity, <i>n</i> (%) | |
| White | 208 (40.2) |
| Black | 211 (40.8) |
| Hispanic or Latino | 28 (5.4) |
| Other | 52 (10.1) |
| Unknown | 18 (3.5) |
| Seen by cardiologist during the past year, <i>n</i> (%) | 309 (59.8) |
| Previous myocardial infarction, <i>n</i> (%) | 71 (13.7) |
| Coronary artery disease, <i>n</i> (%) | 265 (51.3) |
| Hypertension, <i>n</i> (%) | 411 (79.5) |
| Diabetes, <i>n</i> (%) | 199 (38.5) |
| Valvular heart disease, <i>n</i> (%) | 91 (17.6) |
| Atrial fibrillation, <i>n</i> (%) | 117 (22.6) |
| Asthma, <i>n</i> (%) | 96 (18.6) |
| COPD, <i>n</i> (%) | 55 (10.6) |

* There were 517 patients included in the study. All patient characteristics were determined from data stored in discrete fields within the electronic health record. See the Appendix (available at www.annals.org) for the diagnostic codes used to identify inclusion criteria and chronic diseases. COPD = chronic obstructive pulmonary disease; GIM = general internal medicine.

intake history, medical history, or notes appended to the problem list; if an LVEF was recorded, the value was recorded and was used to determine eligibility for the β -blocker and ACE inhibitor or ARB measure. This review was facilitated by an electronic text search by using the terms “lvef,” “ef_,” “echo,” “cardiac,” and “ventricular” to identify any physician note that contained 1 of these terms. Qualitative descriptions were assigned to the following LVEF values: “normal,” 55%; “mildly reduced,” 45%; “moderately reduced,” 35%; and “severely reduced,” 25%. All patients who met the eligibility criteria for heart failure were included in the denominator of this quality measure.

Medication Quality Measures: Automated Electronic Health Record Review

The automated assessment for the medication quality measures involved a series of stepwise queries. We first identified which eligible patients had diagnoses of heart failure in their encounter, problem list, or medical history (Appendix, available at www.annals.org). Next, we queried the patient encounter data table, the problem list data table, and the medical history data table for all diagnoses (ICD-9 codes) that were listed in the specifications as exclusion criteria for the quality measures (Appendix, available at www.annals.org). Finally, the medication table was queried for current and discontinued prescriptions for β -blockers, ACE inhibitors, ARBs, and warfarin. A medi-

cation was considered current if the start date fell before or within the study interval and the end date was either after the interval or absent. A medication was considered discontinued only if an end date was entered before the end of the study interval (31 December 2004). The medication list through 2004 included “historical” medications that were recorded in the EHR but were prescribed by a doctor in another practice. All data were then merged by using the unique patient identifier in the EHR.

Medication Quality Measures: Manual Electronic Health Record Review

For each of the 3 medication measures, we identified the subset of patients who did not meet the quality measure. An investigator then conducted a structured manual review of the data in the EHR, including physician notes, for all such patients. This method is often called a “hybrid” review (12). The review was facilitated by an available free-text search function (for example, find “ACE”) that allowed us to identify information in notes that would otherwise be difficult to find. The review focused on patients who did not have an indicated medication based on the automated review to determine whether the medication or an exclusion criterion was described in a clinician’s notes. We also reviewed the charts of patients who were not prescribed a β -blocker, an ACE inhibitor, or an ARB and who met exclusion criteria on the basis of the specified diagnoses for bradycardia or hypotension to determine whether these patients had valid reasons for not starting or for discontinuing these medications.

Statistical Analysis

Performance on the medication quality measures was calculated by using the formula recommended by the PCPI: N_1 prescribed / (N_1 prescribed + N_2 not prescribed and no exclusion). We calculated the sensitivity of automated searches of the EHR for identifying patients who were receiving a recommended medication and those who had a valid exclusion criterion relative to automated searches that were supplemented by manual review of electronic charts. If a medication was recorded in the chart, we assumed that a prescription had actually been written and that the patient was instructed to take the medication; thus, we did not assess the number of false-positive documented medications. We could only assess the false-positive rates for a limited number of exclusion criteria, such as bradycardia or hypotension. For other exclusion criteria (for example, aortic stenosis for ACE inhibitors or asthma for β -blockers), we had no way of disproving that the condition existed.

Role of the Funding Source

This study was funded by the Agency for Healthcare Research and Quality (AHRQ). The AHRQ had no role in the design, conduct, or analysis and was not involved in the decision to submit the manuscript for publication.

RESULTS

Five hundred seventeen patients met the eligibility criteria (Table 1). Most patients had codes for heart failure recorded in 2 or more locations (that is, encounter diagnosis, problem list, or medical history). The average number of visits over the previous 18 months was 6.4 (SD, 14.4); 51.3% of patients had underlying coronary artery disease, and 79.5% had a history of hypertension.

Validation of Heart Failure Diagnosis

Twenty-eight patients had a diagnosis of heart failure on their problem list or medical history but did not have heart failure listed as an encounter diagnosis, which indicated that physicians had never used heart failure as the reason for a visit or as a reason for ordering a test, medication, or referral at the time of a clinic visit. Of these patients, 12 (42.9%) had documentation of clinical heart failure in physicians' notes, 7 (25%) had asymptomatic left ventricular dysfunction with LVEF less than 0.45, 8 (28.6%) had a normal LVEF, and 1 (3.6%) did not have an LVEF recorded. Of the 66 patients who had only a single-encounter diagnosis of heart failure in the EHR, 34 (51.5%) had clinical heart failure described and addressed in physician notes, 5 (7.6%) had a "history of heart failure" when initially seen but did not have evidence of active heart failure during the study period, and 3 (4.5%) patients had a brief episode of high-output failure due to severe anemia. Eleven (16.7%) had definitive evidence that they did not have clinical heart failure; in these cases, the diagnosis code had clearly been used to rule out heart failure (for example, when referring a patient for echocardiography). In 13 patients (19.7%), documentation was inadequate for a decision to be made. Thus, the automated review identified 40 patients who seemed not to have heart failure. Exclusion of these 40 patients from all subsequent analyses did not change the results, and we only report the rates for the overall population. Under the assumption that all patients with multiple heart failure encounter diagnoses truly had heart failure (these cases were not reviewed), the automated review correctly identified 477 of 517 patients with heart failure (92.3%).

Measurement of Left Ventricular Ejection Fraction

Five hundred three patients (97.3%) had documentation of LVEF measurement, 489 (94.6%) were identified by electronic search, and 14 (2.7%) were identified by additional manual review of physician notes in the EHR (Table 2). Most patients with tests recorded in the EHR had multiple LVEF measurements, which caused discordance in classifying patients as having LVEF less than 0.40 or 0.40 or greater (Table 2); 150 (29.8%) patients had 1 or more LVEF measurement less than 0.40 and 1 or more LVEF measurement 0.40 or greater. The most common discordant pattern was a first LVEF less than 0.40 and the most recent LVEF was 0.40 or greater.

For the medication performance measures that were only applicable to patients with LVEF less than 0.40 (pre-

Table 2. Left Ventricular Ejection Fraction Measurement*

| Variable | Patients, n (%) |
|---|-----------------|
| Total with documentation of LVEF measurement | 503 (97.3) |
| LVEF found by electronic search | 489 (94.6) |
| LVEF found by review of physician notes | 14 (2.7) |
| No recorded LVEF | 14 (2.7) |
| LVEF measurements in EHR | |
| 0 | 28 (5.4) |
| 1 | 70 (13.5) |
| 2 | 106 (20.5) |
| 3 | 101 (19.5) |
| ≥4 | 212 (41.0) |
| Individual variations in LVEF measurements† | |
| All values < 0.40 | 104 (20.7) |
| All values ≥ 0.40 | 249 (49.5) |
| One or more values < 0.40 and 1 or more values ≥ 0.40 | 150 (29.8) |
| First < 0.40 and latest ≥ 0.40 | 75 (14.9) |
| First ≥ 0.40 and latest < 0.40 | 33 (6.6) |
| Other patterns of variation | 42 (8.3) |

* EHR = electronic health record; LVEF = left ventricular ejection fraction.

† A total of 288 LVEF measurements in 195 patients had only a qualitative description of their LVEF; 13 patients had only a qualitative description of their LVEF and no study with a quantitative estimate of the LVEF. If only a qualitative description of the LVEF was available from a study, patients described as having "moderately" or "severely" depressed LVEF were assigned to LVEF < 0.40 and those described as having "mildly" depressed LVEF were assigned to LVEF ≥ 0.40. Percentages shown are based on a denominator of 503 patients with at least 1 LVEF measurement available. If the sample is restricted to the 419 patients with 2 or more LVEF measurements, then the percentage of patients with 1 or more values < 0.40 and 1 or more values ≥ 0.40 was 35.8%.

scription of ACE inhibitor, ARB, or β -blocker), we defined patients as eligible if they had any LVEF less than 0.40 ($n = 254$; 49.1% of all patients).

Prescription of a β -Blocker for Patients with Systolic Dysfunction

Of the 254 patients with any LVEF less than 40%, automated review of EHR detected 219 (86.2%) patients with an active electronic prescription for a β -blocker. Of the remaining 35 patients, 13 met 1 or more exclusion criteria: asthma or chronic obstructive pulmonary disease ($n = 8$), heart rate less than 50 beats per minute on 2 consecutive visits ($n = 5$), or a diagnosis of bradyarrhythmia ($n = 2$). Thus, performance on the β -blocker quality measure was 90.9% by using automated EHR review (Table 3).

Manual review of clinicians' notes in the EHR did not reveal any additional patients who were being treated with a β -blocker. However, we identified 6 additional patients who met the exclusion criteria: β -blocker was listed under "allergies" for 2 patients, 2 patients had symptomatic bradycardia based on more lenient conditions than were specified (heart rate < 50 beats/min on any 2 previous visits) and had documentation in physician notes that the bradycardia resulted from β blockade, 1 patient declined to take a β -blocker, and 1 patient could not afford the medication. Thus, 6 of 22 patients (27.3%) without an active prescription for a β -blocker and no exclusion criteria detected by automated EHR review had a valid exclusion documented and detected by manual review; they were reclassified as

Table 3. Performance of Automated Review Compared with Automated Followed by Manual Review*

| Measure | Patients, n | Automated Review | | | | Changes with Manual Review of Physician Notes | | | Automated Followed by Manual Review | | | |
|--|-------------|--------------------------|----------------------------|--|--|---|---------------------------------|---|-------------------------------------|----------------------------|--|--|
| | | Prescribed Medication, n | Met Exclusion Criteria, nt | No Exclusion Criteria and Not Prescribed Medication, n | Eligible and Prescribed Medication, n/n (%)† | Detected Prescribed Medication, n | Detected Exclusion Criteria, nt | Identified False-Positive Exclusion Criteria, n | Prescribed Medication, n | Met Exclusion Criteria, nt | No Exclusion Criteria and Not Prescribed Medication, n | Eligible and Prescribed Medication, n/n (%)‡ |
| β-Blocker for LVEF values < 0.40§ | 254 | 219 | 13 | 22 | 219/241 (90.9) | 0 | 6 | 1 | 219 | 18 | 17 | 219/236 (92.8) |
| ACE inhibitor or ARB for LVEF values < 0.40§ | 254 | 217 | 23 | 14 | 217/231 (93.9) | 5 | 8 | 2 | 222 | 29 | 3 | 222/225 (98.7) |
| Warfarin for comorbid atrial fibrillation | 117 | 69 | 19 | 29 | 69/98 (70.4) | 4¶ | 20 | 0 | 73 | 39 | 5 | 73/78 (93.6) |

* ACE = angiotensin-converting enzyme; ARB = angiotensin-receptor blocker; LVEF = left ventricular ejection fraction.
 † The exclusion criteria for the automated measurements included only contraindications that could be determined by using diagnosis codes, allergies, or other information stored in unique fields. For the chart review, exclusion criteria also included documentation in physicians' notes of contraindications, adverse events from medications, and patient reasons for not prescribing medications (e.g., patient declined to take medication, inability to afford a medication, or persistent nonadherence to a medication). See the **Appendix** (available at www.annals.org) for the diagnostic codes for the exclusion criteria used in the automated measurement.
 ‡ The percentage eligible to be prescribed a medication was calculated as the number of patients prescribed medication/(number prescribed medication + number not prescribed measure and no exclusion criteria).
 § Patients were considered eligible for the β-blocker measure and the ACE inhibitor/ARB measure if all LVEFs recorded were < 0.40 (n = 104) or if they had ever had an LVEF < 0.40 (n = 150).
 || 170 patients were prescribed an ACE inhibitor and 51 were prescribed an ARB; 4 patients were prescribed both.
 ¶ 2 patients were prescribed warfarin on the basis of chart review, and 2 were prescribed low-molecular-weight heparin, which was not included in the original measure specifications.

“met exclusion criteria.” However, 1 patient who met the exclusion criterion for hypotension on automated EHR review actually had orthostatic hypotension that was not secondary to β-blocker administration; this patient was reclassified to “no exclusion, no prescribed medication.” Thus, performance on the β-blocker quality measure was 92.8% by using automated and manual EHR review (Table 3).

The sensitivity of the automated quality assessment was 100% for identifying patients with heart failure who were taking a β-blocker (Table 4). However, the automated quality assessment detected only 12 of 18 patients with valid exclusion criteria (66.7%), and 1 of 13 patients who met an exclusion criterion was judged not to have a true exclusion criterion (false-positive rate, 7.7%).

Prescription of an Angiotensin-Converting Enzyme Inhibitor or Angiotensin-Receptor Blocker for Patients with Systolic Dysfunction

Of the 254 patients with any LVEF less than 0.40, automated EHR review detected 217 (85.4%) with an active electronic prescription for an ACE inhibitor or ARB (Table 3). Of the remaining 37 patients, 23 (62.2%) met 1 or more of the exclusion criteria: end-stage renal disease (n = 13), aortic stenosis (n = 5), “allergy” or intolerance (n = 7), and hypotension (n = 2). Thus, performance on the ACE inhibitor and ARB quality measure was 93.9% by using automated EHR review (Table 3).

Among the 14 patients without an active prescription for an ACE inhibitor or an ARB in the EHR, manual review of clinicians' notes in the EHR revealed that 5 patients had been prescribed an ACE inhibitor or ARB that

was not recorded in the medication list; these patients were reclassified from “no exclusion, not prescribed medication” to “prescribed medication.” Six patients were found on chart review to have exclusion criteria (hypotension, dizziness, renal insufficiency, hyperkalemia, and persistent nonadherence); these patients were reclassified from “no exclusion, not prescribed medication” to “met exclusion criteria.” In addition, 2 patients met the exclusion criterion for hypotension on automated review, but chart review found that the hypotension was not related to an ACE inhibitor or ARB; these patients would have been reclassified as “no exclusion, not prescribed medication,” but they met other exclusion criteria on chart review and therefore remained classified as “met exclusion criteria.” Thus, only 3 patients who were not prescribed an ACE inhibitor or ARB had no exclusion criteria, and performance on the ACE inhibitor or ARB quality measure was 98.7% by using automated and manual EHR review (Table 3).

The automated quality assessment had a sensitivity of 97.7% for identifying patients with heart failure taking an ACE inhibitor or ARB (Table 4). However, the automated quality assessment captured only 21 of 29 patients with valid exclusion criteria (sensitivity, 72.4%), and 2 of 23 patients who met exclusion criteria were judged not to have a true exclusion (false-positive rate, 8.7%).

Anticoagulation for Patients with Heart Failure and Atrial Fibrillation

Of the 117 patients with heart failure who also had a diagnosis of atrial fibrillation, automated EHR review detected 69 (59.0%) with an active electronic prescription for

warfarin. Of the remaining 48 patients, 19 (39.6%) met 1 or more exclusion criteria. Most had a history of bleeding complications, and a smaller number had characteristics that placed them at high risk for bleeding complications (for example, liver disease or mental disorder). Thus, performance on the warfarin quality measure was 70.4% using automated EHR review (Table 3).

Among the 29 patients who did not have an active prescription for warfarin in the EHR and did not meet the previously defined exclusion criteria, manual review of clinicians' notes in the EHR revealed that 2 patients had warfarin use documented in the physician notes and 2 were using low-molecular-weight heparin for anticoagulation. Of the other 25 patients not receiving anticoagulation, 5 had paroxysmal atrial fibrillation with a clear, transient precipitant (for example, coronary artery bypass surgery), and the physician had documented that it was thought unlikely to recur and that long-term anticoagulation was not considered necessary; 3 patients had ablation therapy for paroxysmal atrial fibrillation that was thought to have been curative (that is, no documented recurrence since the time of the procedures). Another 4 patients had diagnosis codes in their EHR indicating bleeding complications (2 had severe epistaxis and 2 had diverticulosis with bleeding) that were not included in the original list of exclusionary criteria, 3 had bleeding complications documented in physician notes, 4 had a history of falls with documentation that the physician thought the risks of warfarin were greater than the benefits, and 1 patient declined warfarin therapy. Thus, performance on the warfarin quality measure was 93.6% by using automated and manual EHR review (Table 3).

The sensitivity of the automated quality assessment for

identifying patients with heart failure and atrial fibrillation who were taking warfarin was 94.5% (Table 4). However, the automated quality assessment captured only 19 of 39 patients with valid exclusion criteria (sensitivity, 48.7%). All patients with exclusion criteria on automated review were judged to have valid exclusion criteria based on the face validity of the diagnosis codes.

DISCUSSION

To our knowledge, this study is the first to analyze the accuracy of the specifications designed by CMS and PCPI (17) to measure the outpatient quality of care for heart failure according to the ACC-AHA-PCPI measures (15–17). Our results show the challenges that must be overcome to develop automated systems to measure quality with a high degree of accuracy by using only EHR data. First, defining the eligible patient population was problematic. For example, some patients who met the inclusion criteria for a diagnosis of heart failure in their medical history, problem list, or encounter diagnoses actually had asymptomatic left ventricular systolic dysfunction or heart failure that had resolved after definitive treatment of atrial fibrillation (ablation therapy). In addition, most patients in our study had multiple LVEF measurements, and approximately one third had 1 value less than 0.40 and another value 0.40 or greater. We used any LVEF less than 0.40 as a criterion for eligibility for the β -blocker and ACE inhibitor and ARB measures, but national guideline groups need to define a uniform method for defining eligibility for these measures. The difficulty of clearly identifying patients who

Table 4. Sensitivity of Automated Methods for Detecting Qualifying Medications and Exclusion Criteria and Rate of False-Positive Exclusions Using Automated Methods Compared with Review of Physician Notes*

| Variable | Sensitivity of Automated Methods for Detecting Qualifying Medications, n/n (%) [95% CI]† | Sensitivity of Automated Methods for Detecting Exclusion Criteria, n/n (%) [95% CI]‡ | False-Positive Exclusions Using Automated Methods, n/n (%) [95% CI]§ |
|---|--|--|--|
| β -Blocker for LVEF < 0.40† | 219/(219 + 0) 219/219 (100) [98.3–100.0] | (13 – 1)/(12 + 6) = 12/18 12/18 (66.7) [41–86.6] | 1/13 1/13 (7.7) [0.2–36.0] |
| ACE inhibitor or ARB for LVEF < 0.40 | 217/(217 + 5) 217/222 (97.7) [94.8–99.3] | (23 – 2)/(21 + 8) = 21/29 21/29 (72.4) [52.8–87.3] | 2/23 2/23 (8.7) [1.1–28.0] |
| Warfarin for comorbid atrial fibrillation | 69/(69 + 4) 69/73 (94.5) [86.6–98.5] | (19 – 0)/(19 + 20) = 19/39 19/39 (48.7) [32.4–65.2] | 0/19 0/19 (0) [0.0–17.6] |

* ACE = angiotensin-converting enzyme; ARB = angiotensin-receptor blocker; LVEF = left ventricular ejection fraction.

† Sensitivity is defined as the number of patients with a medication identified by the automated review (true positives) divided by the number of true positives plus the number of patients with a medication identified by manual review of providers' electronic notes if the medication was not identified by the automated review (false negatives). If a medication was shown as being electronically prescribed, we assumed that a prescription had actually been signed and the patient had been instructed to take the medication. Therefore, rate of false positives for medications was not assessed.

‡ Sensitivity is defined as the number of patients with an exclusion criterion identified by the automated review that was subsequently validated on hybrid review (true positives) divided by the number of true positives plus the number of patients with exclusion criteria based upon hybrid review that were not identified by the automated review (false negatives).

§ The rate of false-positive exclusion criteria was defined as the number of patients with an exclusion criterion identified by the automated review that was found to be invalid on review of physicians' notes (false positives) divided by the total number of patients identified as having an exclusion criterion by the automated review (Table 3).

|| Patients were considered eligible for the β -blocker measure and the ACE inhibitor/ARB measure if all LVEFs recorded were < 0.40 ($n = 104$) or if they had ever had an LVEF < 0.40 ($n = 150$).

are eligible for a quality measure will probably arise for other conditions and preventive services.

Although automated review of EHR was highly sensitive for identifying patients' current medications (Table 3), sensitivity for detecting exclusion criteria was more problematic, ranging from 48.7% to 72.4% (Table 4). As a result, performance on the medication quality measures was lower with automated review than with automated plus manual EHR review. We attempted to improve the ability to detect patients with valid exclusion criteria by identifying patients who had been prescribed an indicated medication in the past (that is, by searching the EHR for past medications) but did not have an active prescription at the end of the study period. However, only half of the patients who had previously been prescribed a β -blocker, an ACE inhibitor, or an ARB had a documented reason for discontinuing the medication (data not shown). Some patients had their medications withdrawn for no apparent reason around the time of hospital admission and discharge, which suggests that the discontinuation was actually a medication reconciliation error (18, 19). Therefore, if a patient previously was prescribed a medication but no longer has an active prescription in the EHR, one cannot assume the presence of a valid exclusion criterion. Rather, this should be cause for investigating a possible inadvertent omission (that is, medication reconciliation error).

If we are to improve the accuracy of automated measures of quality for the care of patients with heart failure, clinicians will need to improve their documentation of exclusion criteria in the medical history or problem list. Having more accurate information will help physicians and health care systems identify ideal candidates for a therapy who currently are not receiving it, including patients who have had medication reconciliation errors that caused inadvertent discontinuation of treatment. Conversely, if a patient's usual physician documents the reason for not prescribing a medication, then this should help prevent other physicians from prescribing a medication that may be harmful. Thus, improving documentation in EHR systems should allow physicians to more proactively identify problems and improve quality and safety.

Our study has several limitations. We only conducted the study in a single practice using a single EHR. We do not know the generalizability of our findings to other sites with other EHR systems or to sites with less experience using an EHR. We did not assess the PCPI measures of blood pressure and weight monitoring. Automated review is likely to perform very well for these measures, so providers' overall performances probably will be better than those based only on assessment of medications. In addition, the overall agreement between automated review and hybrid review probably would have been higher if these measures were included. We also could not perform automated reviews to assess performance for the measures that addressed clinical symptoms and signs of volume overload and assessment of patients' activity levels. It may be desirable to

capture this information in discrete fields for quality measurement, but this may slow physician workflow and may be unacceptable.

Because we used EHR data to identify eligible patients, it is possible that some patients were not included because their heart failure was documented only in physician notes. Similarly, because we used EHR data to identify patients with heart failure who had a diagnostic test that routinely measures LVEF, it is possible that some patients who had LVEF measured were not included because documentation of the test result could be found only in physician notes. We could not assess false-positive documentation of medications and of all exclusion criteria. Finally, we did not manually review physicians' notes for every patient, so we could not determine the true sensitivity and specificity of automated review according to the standard definitions of these terms.

Nevertheless, our findings have implications for national organizations that develop quality measures, EHR vendors, provider organizations interested in using EHRs for quality improvement, and payers who would like to use these data in pay-for-performance programs. First, diagnosis codes available in EHRs should be designed to facilitate quality-of-care measurement. For example, a diagnosis of "asymptomatic left ventricular systolic dysfunction" should be created so physicians are not forced to use a diagnosis of heart failure as a surrogate. Current EHR systems mostly rely on ICD-9 codes, which are poorly suited for use in clinical care or quality measurement. The more recently developed Systemized Nomenclature of Medicine (SNOMED) terminology provides a common language that enables a more consistent way of capturing, sharing, and aggregating health data across specialties and sites of care (20). However, the terms are not specifically designed to facilitate quality measurement.

Experts developing guidelines and quality measures should recognize that most patients with heart failure will have multiple measures of LVEF that may vary over time, and guideline recommendations and quality measures should reflect this reality. Because ejection fraction may improve markedly with current therapies or decline over time with new or persistent myocardial injury, it may be helpful to create diagnostic codes that clearly classify the patients' status (for example, "heart failure with systolic dysfunction"). Electronic health record vendors should always include fields that can be used to document LVEF and identify patients for whom ACE inhibitors, ARBs, and β -blockers are indicated.

Diagnostic codes that allow physicians to easily and accurately capture exclusion criteria should be included within EHRs. The new Current Procedural Terminology category II codes that enable documentation of medical or patient reasons for excluding a patient from a performance measure may facilitate a standard approach to this (21). Capturing patient-related reasons for nonadherence is particularly important so that providers who care for vulnera-

ble populations do not seem to be delivering worse quality of care because they have more patients with financial barriers (22, 23). Capturing exclusion criteria would be best accomplished if EHR vendors create “quality dashboards” in which physicians could simply select a reason from a list to document why they are not ordering a recommended treatment or diagnostic test. This functionality would improve the accuracy of quality measurement and aid clinical care by allowing physicians to track the reason a patient did not receive a recommended treatment, allow this decision to be revisited if circumstances change (for example, when a treatment becomes more affordable or a contraindication is resolved), and more accurately identify patients who truly need to have a therapy initiated for point-of-care alerts or outreach programs.

For provider organizations, our results show that if electronic alerts delivered to providers at the point of care are only based on the specifications used in this study, they will often be incorrect. More accurate electronic capture of critical information for quality measurement can be used to improve current reminder systems and disease management programs and simultaneously allow providers to objectively demonstrate the quality of their care. Finally, until improvements are made in the ability to accurately measure quality of care using only discrete data within EHRs (as opposed to also reviewing physician notes), this type of quality measure should be used cautiously in pay-for-performance programs. Most patients who appeared to “fail” the quality indicator at this high-performing site were incorrectly classified. When actual quality of care is high and capture of exclusion criteria is only fair, providers may be judged by how well they are able to document exclusion criteria rather than on the basis of the true quality of their care (10).

From Northwestern University, Chicago, Illinois.

Acknowledgment: The authors thank Heidi Bossley for technical assistance.

Grant Support: This project was supported in part by grant number 5 U18 HS013690 from the Agency for Healthcare Research and Quality.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: David Baker, MD, MPH, 676 North St. Clair Street, Suite 200, Chicago, IL 60611-2927; e-mail, dwbaker@northwestern.edu.

Current author addresses are available at www.annals.org.

References

1. Institute of Medicine. Leadership by Example: Coordinating Government Roles in Improving Health Care Quality. Washington, DC: National Acad Press; 2002.
2. Epstein AM, Lee TH, Hamel MB. Paying physicians for high-quality care.

- N Engl J Med. 2004;350:406-10. [PMID: 14736934]
3. Medicare Payment Advisory Commission. Report to the Congress: Variation and Innovation in Medicare. Washington, DC: Medicare Payment Advisory Committee; 2003.
4. Milgate K, Cheng SB. Pay-for-performance: the MedPAC perspective. Health Aff (Millwood). 2006;25:413-9. [PMID: 16522581]
5. Centers for Medicare & Medicaid Services. Quality initiatives—general information. Accessed at www.cms.hhs.gov/QualityInitiativesGenInfo/ on 18 August 2006.
6. Jha AK, Li Z, Orav EJ, Epstein AM. Care in U.S. hospitals—the Hospital Quality Alliance program. N Engl J Med. 2005;353:265-74. [PMID: 16034012]
7. Centers for Medicare & Medicaid Services. Physician Focused Quality Initiative. Accessed at www.cms.hhs.gov/physicianfocusedqualinit/ on 18 August 2006.
8. Boyd CM, Darer J, Boulton C, Fried LP, Boulton L, Wu AW. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. JAMA. 2005;294:716-24. [PMID: 16091574]
9. Tinetti ME, Bogardus ST Jr, Agostini JV. Potential pitfalls of disease-specific guidelines for patients with multiple conditions. N Engl J Med. 2004;351:2870-4. [PMID: 15625341]
10. Doran T, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U, et al. Pay-for-performance programs in family practices in the United Kingdom. N Engl J Med. 2006;355:375-84. [PMID: 16870916]
11. Walter LC, Davidowitz NP, Heineken PA, Covinsky KE. Pitfalls of converting practice guidelines into quality measures: lessons learned from a VA performance measure. JAMA. 2004;291:2466-70. [PMID: 15161897]
12. Kerr EA, Smith DM, Hogan MM, Krein SL, Pogach L, Hofer TP, et al. Comparing clinical automated, medical record, and hybrid data sources for diabetes quality measures. Jt Comm J Qual Improv. 2002;28:555-65. [PMID: 12369158]
13. Borzecki AM, Wong AT, Hickey EC, Ash AS, Berlowitz DR. Can we use automated data to assess quality of hypertension care? Am J Manag Care. 2004;10:473-9. [PMID: 15298233]
14. Persell SD, Wright JM, Thompson JA, Kmetik KS, Baker DW. Assessing the validity of national quality measures for coronary artery disease using an electronic health record. Arch Intern Med. 2006;166:2272-7. [PMID: 17101947]
15. The Physician Consortium for Performance Improvement. Performance measures for heart failure. Accessed at www.ama-assn.org/ama/pub/category/15777.html on 18 August 2006.
16. American College of Cardiology. ACC/AHA clinical performance measures for adults with chronic heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures (Writing Committee to Develop Heart Failure Clinical Performance Measures) endorsed by the Heart Failure Society of America. J Am Coll Cardiol. 2005;46:1144-78. [PMID: 16168305]
17. Doctor's Office Quality—Information Technology (DOQ-IT). Accessed at www.qualitynet.org/dcs/ContentServer?cid=1143577170595&pagename=QnetPublic%2FPage%2FQnetTier2&c=Page On 18 August 2006.
18. Santell JP. Reconciliation failures lead to medication errors. Jt Comm J Qual Patient Saf. 2006;32:225-9. [PMID: 16649654]
19. Parameswaran AC, Tang WH, Francis GS, Gupta R, Young JB. Why do patients fail to receive beta-blockers for chronic heart failure over time? A “real-world” single-center, 2-year follow-up experience of beta-blocker therapy in patients with chronic heart failure. Am Heart J. 2005;149:921-6. [PMID: 15894978]
20. SNOMED. Systemized Nomenclature of Medicine Clinical Terms. Accessed at www.snomed.org/snomedct/index.html on 18 August 2006.
21. American Medical Association. CPT II codes. Accessed at www.ama-assn.org/ama/pub/category/10616.html on 18 August 2006.
22. Franks P, Fiscella K. Effect of patient socioeconomic status on physician profiles for prevention, disease management, and diagnostic testing costs. Med Care. 2002;40:717-24. [PMID: 12187185]
23. Fiscella K, Franks P, Gold MR, Clancy CM. Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care. JAMA. 2000;283:2579-84. [PMID: 10815125]

Current Author Addresses: Drs. Baker, Persell, Soman and Mr. Thompson and Mr. Liss: Feinberg School of Medicine, Northwestern University, 676 North St. Clair Street, Suite 200, Chicago, IL 60611. Dr. Burgner: Northwestern Memorial Hospital, 210 East Huron Road, Chicago, IL 60611. Dr. Kmetik: American Medical Association, 515 State Street, Chicago, IL 60610.

APPENDIX: INTERNATIONAL CLASSIFICATION OF DISEASES, NINTH REVISION, CLINICAL MODIFICATION (ICD-9-CM) CODES USED TO ASSESS ELIGIBILITY, COMORBID CONDITIONS, AND EXCLUSION CRITERIA WITH THE ELECTRONIC HEALTH RECORD

Heart Failure Codes for Study Inclusion

398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, and 428.xx.

Comorbid Conditions

Myocardial infarction: 411.xx and 412; coronary artery disease: 410-410.92, 411-411.89, 412, 413-413.9, 414-414.07, 414.8, 414.9, V45.81, and V45.82; hypertension: 401.0, 401.1, 401.9, 402.xx, 403.xx, and 404.xx; diabetes mellitus: 250-250.93, 357.2, 362.01, 362.02, and 366.41; valvular heart disease: 394.xx, 395.xx, 396.xx, 397.xx, 424.0, and 424.1; atrial fibrillation: 427.3, 427.31, and 427.32; asthma: 493.xx; chronic obstructive pulmonary disease: 491.20, 491.21, 492.0, 492.8, 496, 496.0, 506.4, and 518.2.

Exclusion Criteria for β -Blocker Use

Heart block, second- or third-degree: 426.0, 426.12, 426.13, and 426.7; bradycardia: 427.81, 427.89, and 337.0; hypotension: 458.xx; asthma, chronic obstructive pulmonary disease (see above).

Exclusion Criteria for Angiotensin-Converting Enzyme Inhibitor and/or Angiotensin-Receptor Blocker Use

Renal arterial stenosis: 440.1; end-stage renal disease: V56.0, V56.8, 39.95, 54.98, 788.5, 585, 586, and 584.xx; hypertensive renal disease: 403.01, 403.11, 403.91, 404.02, 404.03, 404.12, 404.13, 404.92, and 404.93; pregnancy: 640-648.94, 650-658.93, 659-659.93, 660-669, 670-677, V22-V22.9, and V23-V23.4; and hypotension: 458.xx.

Exclusion Criteria for Warfarin Use

Intracranial bleed: 430, 431, 432.0, 432.1, 432.9, and 437.3; gastrointestinal bleeding/liver disease: 570, 530.7, 569.3,

571.2, 571.5, 578.0, 578.1, 578.9, 531.00, 531.01, 531.20, 531.21, 531.40, 531.41, 531.60, 531.61, 532.00, 532.01, 532.20, 532.21, 532.40, 532.41, 532.60, 532.61, 533.00, 533.01, 533.20, 533.21, 533.40, 533.41, 533.60, 533.61, 534.00, 534.01, 534.20, 534.21, 534.40, 534.41, 534.60, and 534.61; bleeding from other causes: 280.8, 280.9, 285.1, 459.0, 599.7, 786.3, and 784.7; hematological disorders: 286.0, 286.1, 286.2, 286.3, 286.4, 286.5, 286.6, 286.7, 286.9, 287.3, 287.4, 287.5, 203.00, 203.01, 203.10, 203.11, 203.80, 203.81, 204.00, 204.01, 204.10, 204.11, 204.20, 204.21, 204.80, 204.81, 204.90, 204.91, 205.00, 205.01, 205.10, 205.11, 205.20, 205.21, 205.30, 205.31, 205.80, 205.81, 205.90, 205.91, 206.00, 206.01, 206.10, 206.11, 206.20, 206.21, 206.80, 206.81, 206.90, 206.91, 207.00, 207.01, 207.10, 207.11, 207.20, 207.21, 207.80, 207.81, 208.00, 208.01, 208.10, 208.11, 208.20, 208.21, 208.80, 208.81, 208.90, and 208.91; mental disorders and substance abuse: 291.0, 291.1, 291.2, 291.3, 291.4, 291.5, 291.9, 292.0, 292.2, 292.9, 296.7, 297.0, 297.1, 297.2, 297.3, 297.8, 297.9, 298.0, 298.1, 298.2, 298.3, 298.4, 298.8, 298.9, 291.81, 291.89, 292.11, 292.12, 292.81, 292.82, 292.83, 292.84, 292.89, 295.00, 295.01, 295.02, 295.03, 295.04, 295.05, 295.10, 295.11, 295.12, 295.13, 295.14, 295.15, 295.20, 295.21, 295.22, 295.23, 295.24, 295.25, 295.30, 295.31, 295.32, 295.33, 295.34, 295.35, 295.40, 295.41, 295.42, 295.43, 295.44, 295.45, 295.50, 295.51, 295.52, 295.53, 295.54, 295.55, 295.60, 295.61, 295.62, 295.63, 295.64, 295.65, 295.70, 295.71, 295.72, 295.73, 295.74, 295.75, 295.80, 295.81, 295.82, 295.83, 295.84, 295.85, 295.90, 295.91, 295.92, 295.93, 295.94, 295.95, 296.00, 296.01, 296.02, 296.03, 296.04, 296.05, 296.06, 296.10, 296.11, 296.12, 296.13, 296.14, 296.15, 296.16, 296.20, 296.21, 296.22, 296.23, 296.24, 296.25, 296.26, 296.60, 296.61, 296.62, 296.63, 296.64, 296.65, 296.66, 296.80, 296.81, 296.82, 296.89, 296.90, and 296.99.

An "xx" after the decimal points indicates that all codes with the 3 digits shown before the decimal point were used regardless of the latter 2 digits.

Exclusion criteria were only assessed for patients who did not have a medication prescribed. Thus, if a patient was prescribed a medication but had an exclusion criterion, the patient was included in the numerator and the denominator of the performance measure. If a patient was not prescribed a medication and met 1 or more of the exclusion criteria, the patient was removed from the numerator and the denominator.