Annals of Internal Medicine | ORIGINAL RESEARCH

# Accuracy of Electronically Reported "Meaningful Use" Clinical Quality Measures

## A Cross-sectional Study

Lisa M. Kern, MD, MPH; Sameer Malhotra, MD, MA; Yolanda Barrón, MS; Jill Quaresimo, RN, JD; Rina Dhopeshwarkar, MPH; Michelle Pichardo, MPH; Alison M. Edwards, MStat; and Rainu Kaushal, MD, MPH

**Background:** The federal Electronic Health Record Incentive Program requires electronic reporting of quality from electronic health records, beginning in 2014. Whether electronic reports of quality are accurate is unclear.

**Objective:** To measure the accuracy of electronic reporting compared with manual review.

**Design:** Cross-sectional study.

**Setting:** A federally qualified health center with a commercially available electronic health record.

**Patients:** All adult patients eligible in 2008 for 12 quality measures (using 8 unique denominators) were identified electronically. One hundred fifty patients were randomly sampled per denominator, yielding 1154 unique patients.

**Measurements:** Receipt of recommended care, assessed by both electronic reporting and manual review. Sensitivity, specificity, positive and negative predictive values, positive and negative likelihood ratios, and absolute rates of recommended care were measured.

**Results:** Sensitivity of electronic reporting ranged from 46% to 98% per measure. Specificity ranged from 62% to 97%, positive predictive value from 57% to 97%, and negative predictive value from 32% to 99%. Positive likelihood ratios ranged from 2.34 to 24.25 and negative likelihood ratios from 0.02 to 0.61. Differences between electronic reporting and manual review were statistically significant for 3 measures: Electronic reporting underestimated the absolute rate of recommended care for 2 measures (appropriate asthma medication [38% vs. 77%; $P < 0.001$] and pneumococcal vaccination [27% vs. 48%; $P < 0.001$]) and overestimated care for 1 measure (cholesterol control in patients with diabetes [57% vs. 37%; $P = 0.001$]).

**Limitation:** This study addresses the accuracy of the measure numerator only.

**Conclusion:** Wide measure-by-measure variation in accuracy threatens the validity of electronic reporting. If variation is not addressed, financial incentives intended to reward high quality may not be given to the highest-quality providers.

**Primary Funding Source:** Agency for Healthcare Research and Quality.

*Ann Intern Med.* 2013;158:77-83. www.annals.org
For author affiliations, see end of text.

The U.S. government has launched an unprecedented program to promote "meaningful use" of electronic health records (EHRs) (1). This program is based on the premise that it is not sufficient for providers to adopt EHRs; rather, they should actively use them to track and improve quality (2). The Electronic Health Record Incentive Program offers up to $27 billion in incentives for meaningful use beginning in 2011 (2). Those who do not achieve meaningful use by 2015 face financial penalties (3). A core objective of the program is that providers report "clinical quality measures" to the Centers for Medicare & Medicaid Services or the states (4). Providers will initially attest to their quality performance, but by 2014, they are expected to submit measures from their EHRs (5).

Historically, quality measures have been derived from administrative claims or manual review of paper records (6). Although administrative claims can generate data for large samples of patients, they lack clinical detail. Manual review can generate clinical detail but only on small samples of patients because of its time-consuming nature. Automated electronic reports of quality from EHRs can potentially address these limitations, offering clinically detailed data for many patients.

However, whether electronically reported measures are valid representations of delivered care is unclear. Compared with manual review, the current reference standard, electronically reported measures could underestimate quality (for example, if documentation of provision of recommended care resides primarily in free-text progress notes) or overestimate quality (for example, if the process captures tests "ordered" when the specifications call for tests "completed"). Previous studies have reported differences in health care quality measured by electronic reporting and manual review (7–12), but most studies addressed quality for only 1 clinical condition, and most were done in academic settings or integrated delivery systems (7–9, 11, 12).

We sought to test the accuracy of electronic reporting in a community-based setting for 12 quality measures, 11 of which are represented in Stage 1 Meaningful Use Clinical Quality Measures. If electronic reporting is not accu-

*See also:*

**Print**
Editorial comment. . . . . 131

 

***Context***

The U.S. government's meaningful use program incentivizes providers to use electronic health records to improve health care quality.

***Contribution***

This study compared electronic and manual chart documentation for 12 process-of-care measures in a single medical system and found that the accuracy of electronic reporting was highly variable. Electronic reporting overestimated provision of some measures and underestimated others.

***Caution***

The study was conducted before formal launch of the meaningful use program.

***Implication***

The accuracy of electronic reporting of quality measures is highly variable. Electronic health records must provide platforms that facilitate accurate reporting by providers if they are to be an important way of improving health care quality.

*—The Editors*

rate, financial incentives intended to reward high-quality care may not be given to the highest-quality providers.

## METHODS

### Overview

The study is a cross-sectional comparison of care quality at the practice level assessed by automated electronic reporting and manual review of electronic records, done in 2010 using 2008 data. It replicates many definitions and characteristics of the meaningful use program, but clinical providers were not receiving financial bonuses for proper documentation and achieving meaningful use or other internal benchmarks at the time care was delivered.

The institutional review boards of Weill Cornell Medical College and the practice approved the protocol.

### Setting

Data on care quality came from a federally qualified health center (FQHC), the Institute for Family Health (IFH), which serves 75 000 patients making 225 000 visits at 6 sites each year. Approximately 25% of patients are black or Hispanic (75% are white) and 50% earn incomes below the federal poverty level; more than one third receive Medicaid, and approximately 20% are uninsured.

### Context

In 2009, we identified (and a national expert panel validated) a set of 18 quality measures selected for their potential to capture the effects of interoperable EHRs on quality and be reported by automated electronic means (13). This study was designed to validate the electronic reporting of that measure set. When we designed this study, no commercial EHR vendors could report these measures without costly custom programming; therefore, we partnered with IFH, which had implemented a commercially available EHR in 2007 and had its own information technology (IT) staff who had been working independently on automated electronic reporting of 12 of our 18 original quality measures (14). This work includes those 12 measures, 11 of which are among the 44 Stage 1 Meaningful Use Clinical Quality Measures. The 12th measure, whether patients with diabetes had their glucose control measured, was not a meaningful use quality measure but is a widely cited quality measure. Of the 11 measures that overlap with Stage 1 meaningful use, 10 are retained in Stage 2; 1 of the 2 measures of glucose control (hemoglobin $A_{1c}$ level <8%) was not included (5).

As in the meaningful use program, our study used explicit definitions of measure denominators (those eligible for the measure) and numerators (those eligible who received recommended care). Because meaningful use specifications had not yet been released at the time of the study, we used the specifications of the FQHC (**Appendix Table**, available at www.annals.org), which are similar (15) but have variations, for example, in the ages of patients eligible for the measure (aged 51 to 80 years vs. 50 to 75 years for colonoscopy screening) and in the target value for disease control (7% vs. 8% for hemoglobin $A_{1c}$ level).

In Stages 1 and 2 of the meaningful use program, providers are required to meet all "core" measures, one of which is to report selected clinical quality measures electronically. Although some core measures have specific targets for minimum performance, clinical quality measures do not have such targets yet; reporting the actual performance (regardless of the level of performance) currently meets the objective.

Care given by providers external to a health system is counted by the meaningful use program (and by the IFH) as having been provided if it is documented in the EHR. The IFH has an interoperable EHR that shares clinical information across its own multiple sites and with several external laboratories; it incorporates information from other external providers into its EHR through scanning documents and through manual abstraction into structured fields or free text.

### Quality Measures

The study set of 12 quality measures reflects care for 8 populations (denominators) of patients: Those who are eligible for breast cancer, cervical cancer, or colorectal cancer screening or influenza or pneumococcal vaccination or who have asthma, diabetes, or ischemic vascular disease (**Appendix Table**).

### Population and Sample

Among all patients aged 18 years or older who had at least 1 office visit at 1 of the 6 FQHC sites north of New York City in 2008, we searched electronically for patients

who met criteria for the 8 measure denominators, using age; sex; International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), codes; and visits. We then randomly sampled 150 patients from each denominator. We allowed patients to contribute to more than 1 denominator (for example, to be eligible for both diabetes care and colorectal cancer screening).

### Manual Review

Two investigators manually reviewed the EHRs using a structured data abstraction tool based on the specifications of the FQHC. The tool was accompanied by the specifications in full and guidance on where in the chart to look for particular information (structured fields [medical history, problem list, encounter diagnosis, medication list, laboratory results, radiology imaging results, immunization history, and health maintenance], free text, and scanned documents) and what information was considered valid. The tool was prepopulated with the selected patients' medical record numbers and the measures for which they were selected. Reviewers were a physician-informaticist and research nurse who each spent 5 hours in training for this project. They were permitted to discuss data collection and to refine the guidance document. On the basis of a review of the same 40 charts without discussion (20 selected at random for each of 2 denominators: ischemic vascular disease and diabetes), interrater reliability was very high (ischemic vascular disease: cholesterol control [$\kappa = 0.74$], appropriate antithrombotic medication [$\kappa = 0.84$]; diabetes: cholesterol control [$\kappa = 0.89$], hemoglobin $A_{1c}$ test done [$\kappa = 1.00$], hemoglobin $A_{1c}$ level <7% [$\kappa = 1.00$], hemoglobin $A_{1c}$ level >9% or no test [$\kappa = 1.00$]) (16).

### Electronic Reports

The IFH IT staff developed and generated an electronic report based on information from structured fields but not from free-text or scanned documents detailing whether each patient was given recommended care (as defined by the measure numerator). The programming for the electronic report went through a process of quality assurance checking, resolving internal inconsistencies through iterative refinements, before being finalized.

Manual reviewers and IT staff were blinded to each other's ratings.

### Statistical Analysis

We considered manual review as the reference standard and considered patients who met criteria for measure numerators as having received recommended care.

We analyzed the accuracy of electronic reporting using diagnostic test metrics: sensitivity, which was the number of patients who received recommended care according to both electronic reporting and manual review (true positives) divided by the total number of patients who received recommended care by manual review (true positives and false negatives); specificity, which was the number of patients who did not receive recommended care according to both electronic reporting and manual review (true negatives) divided by the total number of patients who did not receive recommended care by manual review (true negatives and false positives); positive predictive value [true positives/(true positives + false positives)]; negative predictive value [true negatives/(true negatives + false negatives)]; positive likelihood ratio [sensitivity/(1 – specificity)]; and negative likelihood ratio [(1 – sensitivity)/specificity].

Then, to assess the direction of any disagreement between electronic reporting and manual review, we calculated the absolute proportion of recommended care for electronic reporting (true positives + false positives)/(total sample) and manual review (true positives + false negatives)/(total sample) and calculated the difference in proportions (electronic reporting minus manual review).

For each statistic, we calculated 95% CIs. We considered a difference in absolute rates of recommended care to be statistically significant if the CI did not cross zero.

Analyses were conducted with SAS, version 9.1 (SAS Institute, Cary, North Carolina), and Excel, version 2007 (Microsoft, Redmond, Washington).

### Role of the Funding Source

This study was funded by the Agency for Healthcare Research and Quality. The funding source played no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, and approval of the manuscript; or decision to submit the manuscript for publication.

## RESULTS

The total numbers of patients eligible for the quality measures at the FQHC in 2008 are shown in **Table 1**. We randomly sampled 150 patients for each denominator and found and included 1154 unique patients. These patients were cared for by 97 providers, including attending physicians, resident physicians, and nurse practitioners. The mean age of patients was 55 years (SD, 19), and they had a median of 4 visits (interquartile range, 2 to 7) in 2008. Nearly two thirds (65%) were women. Patients contributed to a mean of 1.6 quality measures each (SD, 1.1; median, 1.0 [interquartile range, 1.0 to 2.0]).

When we compared electronic reporting with manual review, we determined the numbers of true-positive, false-positive, true-negative, and false-negative results for each measure, as shown in **Table 1**.

Sensitivity and specificity varied considerably by the specific quality measure (**Table 1**). Sensitivity ranged from 46% (for asthma medication) to 98% (for having a hemoglobin $A_{1c}$ test done for patients with diabetes). Specificity ranged from 62% (for cholesterol control in patients with diabetes and for ischemic vascular disease medication) to 97% (for pneumococcal vaccination).

Positive and negative predictive values also varied by measure (**Table 1**). The former varied from 57% (for colorectal cancer screening) to 97% (for having hemoglobin $A_{1c}$ test done for patients with diabetes); the latter varied

*Table 1.* **Validity of Electronically Reported Quality Measures, Compared With Manual Review***

| Measure | Total Patients Eligible, *n* | Sample, *n*† | | | | Sensitivity (95% CI) | Specificity (95% CI) | PPV (95% CI) |
|---|---|---|---|---|---|---|---|---|
| | | TP | FP | TN | FN | | | |
| **Appropriate asthma medication** | 1562 | 53 | 4 | 30 | 63 | 0.46 (0.36–0.55) | 0.88 (0.73–0.97) | 0.93 (0.83–0.98) |
| **Cancer screening** | | | | | | | | |
| Breast cancer | 2658 | 34 | 5 | 96 | 15 | 0.69 (0.55–0.82) | 0.95 (0.89–0.98) | 0.87 (0.73–0.96) |
| Cervical cancer | 8677 | 29 | 5 | 115 | 1 | 0.97 (0.83–1.00) | 0.96 (0.91–0.99) | 0.85 (0.69–0.95) |
| Colorectal cancer | 6691 | 17 | 13 | 106 | 14 | 0.55 (0.36–0.73) | 0.89 (0.82–0.94) | 0.57 (0.37–0.75) |
| **Diabetes** | | | | | | | | |
| Hemoglobin $A_{1c}$ test done | 1867 | 111 | 4 | 33 | 2 | 0.98 (0.94–1.00) | 0.89 (0.75–0.97) | 0.97 (0.91–0.99) |
| Hemoglobin $A_{1c}$ level <7% | 1867 | 61 | 3 | 82 | 4 | 0.94 (0.85–0.98) | 0.96 (0.90–0.99) | 0.95 (0.87–0.99) |
| Hemoglobin $A_{1c}$ level >9% or no test | 1867 | 48 | 4 | 93 | 5 | 0.91 (0.79–0.97) | 0.96 (0.90–0.99) | 0.92 (0.81–0.98) |
| LDL cholesterol level <2.59 mmol/L (<100 mg/dL) | 1867 | 50 | 36 | 58 | 6 | 0.89 (0.78–0.96) | 0.62 (0.51–0.72) | 0.58 (0.47–0.69) |
| **Influenza vaccine, age ≥50 y** | 4502 | 43 | 10 | 95 | 2 | 0.96 (0.85–0.99) | 0.90 (0.83–0.95) | 0.81 (0.68–0.91) |
| **IVD** | | | | | | | | |
| Appropriate antithrombotic medication | 248 | 93 | 20 | 32 | 5 | 0.95 (0.88–0.98) | 0.62 (0.47–0.75) | 0.82 (0.74–0.89) |
| LDL cholesterol level <2.59 mmol/L (<100 mg/dL) | 248 | 59 | 20 | 65 | 6 | 0.91 (0.81–0.97) | 0.76 (0.66–0.85) | 0.75 (0.64–0.84) |
| **Pneumoccocal vaccination** | 3062 | 38 | 2 | 76 | 34 | 0.53 (0.41–0.65) | 0.97 (0.91–1.00) | 0.95 (0.83–0.99) |

FN = false negative; FP = false positive; IVD = ischemic vascular disease; LDL = low-density lipoprotein; LR− = negative likelihood ratio; LR+ = positive likelihood ratio; NPV = negative predictive value; PPV = positive predictive value; TN = true negative; TP = true positive.
* Calculations: sensitivity = TP ÷ (TP + FN); specificity = TN ÷ (TN + FP); PPV = TP ÷ (TP + FP); NPV = TN ÷ (TN + FN); LR+ = sensitivity ÷ (1 − specificity); and LR− = (1 − sensitivity) ÷ specificity.
† 150 total patients per measure.

from 32% (for asthma medication) to 99% (for cervical cancer screening).

Positive and negative likelihood ratios varied as well (**Table 1**). Positive likelihood ratios varied from 2.34 (for cholesterol control in patients with diabetes) to 24.25 (for cervical cancer screening). Negative likelihood ratios varied from 0.02 (for having a hemoglobin $A_{1c}$ test done for patients with diabetes) to 0.61 (for asthma medication).

When we considered the absolute rates of recommended care (**Table 2**), 3 measures with electronic reporting–manual review differences were statistically significant. Electronic reporting underestimated the rates of appropriate asthma medication (absolute difference, −39% [95% CI, −50% to −29%]; $P < 0.001$) and pneumococcal vaccination (absolute difference, −21% [CI, −32% to −11%]; $P < 0.001$) compared with manual review. It overestimated the rate of cholesterol control in patients with diabetes (absolute difference, 20% [CI, 9% to 31%]; $P = 0.001$) compared with manual review.

## DISCUSSION

In this study of the accuracy of electronic reporting of quality measures compared with manual review of those measures, we found wide measure-by-measure variation in accuracy and statistically significant differences for 3 measures. Electronic reporting significantly underestimated rates of appropriate asthma medication and pneumococcal vaccination and overestimated rates of cholesterol control in patients with diabetes. There are several possible explanations for these observations. For example, electronic reporting could have underestimated rates of asthma medication and pneumococcal vaccination if care was recorded in free-text notes or scanned documents rather than in structured fields. Electronic reporting could have overestimated rates of cholesterol control if the electronic report and manual reviewers considered different tests to be the "most recent" cholesterol value.

Measure-by-measure variation raises many issues that are integral to quality reporting. For automated reporting to be valid, all of the following must occur: Clinicians have to document the care they deliver and maintain the accuracy of data in EHRs, documentation must be amenable to automated reporting (that is, in structured data fields rather than free text), and electronic specifications have to capture the same fields that a reference standard manual reviewer would consider. There are methods that extract meaning from free text, an approach called "natural language processing" (17, 18); however, they are not available for widespread use.

Previous work found pervasive problems with data accuracy and completeness in the structured fields of EHRs, particularly in problem lists and medication lists (19). Other studies found that evidence of recommended care exists in EHRs but is often only in nonstructured forms, which makes it essentially "missing" from the perspective of automated reporting (20, 21).

Previous studies looked separately at the accuracy of electronic reporting for quality measures related to diabetes

*Table 1*—Continued

| NPV (95% CI) | LR+ (95% CI) | LR− (95% CI) |
|---|---|---|
| 0.32 (0.23–0.43) | 3.83 (1.51–9.73) | 0.61 (0.50–0.76) |
| 0.86 (0.79–0.92) | 13.80 (5.78–32.96) | 0.33 (0.21–0.50) |
| 0.99 (0.95–1.00) | 24.25 (10.07–58.39) | 0.03 (0.00–0.24) |
| 0.88 (0.81–0.93) | 5.00 (2.74–9.13) | 0.51 (0.34–0.75) |
| 0.94 (0.81–0.99) | 8.91 (3.56–22.29) | 0.02 (0.01–0.08) |
| 0.95 (0.89–0.99) | 23.50 (8.28–66.70) | 0.06 (0.02–0.16) |
| 0.95 (0.88–0.98) | 22.75 (8.55–60.53) | 0.09 (0.04–0.22) |
| 0.91 (0.81–0.96) | 2.34 (1.78–3.08) | 0.18 (0.08–0.38) |
| 0.98 (0.93–1.00) | 9.60 (5.39–17.09) | 0.04 (0.01–0.19) |
| 0.86 (0.71–0.95) | 2.50 (1.76–3.55) | 0.08 (0.03–0.20) |
| 0.92 (0.83–0.97) | 3.79 (2.58–5.58) | 0.12 (0.05–0.26) |
| 0.69 (0.60–0.78) | 17.67 (4.91–63.57) | 0.48 (0.38–0.62) |

(8), coronary artery disease (9, 11, 12), and heart failure (7). The magnitude of the discrepancies found between electronic reporting and manual review was substantial. One study found that automated reporting underestimated quality compared with manual review by as much as 15 percentage points for diabetes care (8). A study of treatment with β-blockers after myocardial infarction found that automated reporting, compared with manual review, had a sensitivity of 83% to 100% and a specificity of 17% to 75%, with the variation in accuracy resulting from changes in measure specifications (12). Other studies found that automated reporting underestimated quality of care for coronary artery disease (9, 11) and congestive heart failure (7), primarily due to the failure of automated reporting to fully capture exceptions (that is, medically valid reasons for not prescribing recommended care).

This work adds to the literature by directly comparing, in an underserved population, automated reporting to manual review for 11 meaningful use clinical quality measures. This work includes important findings on the accuracy of measures related to asthma care and pneumococcal vaccination, which were not considered in another recent study (10).

This study has several limitations. First, we identified eligible patients electronically, thus focusing on validating the reporting of numerator data. Validating the denominator was beyond the scope of this study and has been addressed by other investigators, who found that electronic reports that incorporate clinical data from EHRs identify more eligible patients than administrative claims (22). If we corrected for this, our measurements of rates of recommended care (which are already fairly low) would probably be even lower. Second, this study took place at a single FQHC with its own IT staff. We had sufficient statistical power to measure quality at the practice level but not at the provider level. We went through several iterations of quality reporting because we initially detected internal inconsistencies within the data. This type of quality assurance checking would probably not take place in settings without IT support staff and suggests that discrepancies between electronic reporting and manual review in other settings may be even larger than those seen in this study. Third, by defining manual review as the reference standard, this study assumes that electronic reporting cannot outperform manual review and the results support this assumption for now, but as electronic reporting improves, new methods

*Table 2.* **Absolute Rates of Recommended Care, as Measured by Automated Report and Manual Review**

| Measure | Electronic Report | Manual Review | Difference (95% CI) |
|---|---|---|---|
| **Appropriate asthma medication** | 0.38 (0.30 to 0.46) | 0.77 (0.70 to 0.84) | −0.39 (−0.50 to −0.29) |
| **Cancer screening** | | | |
| Breast cancer | 0.26 (0.19 to 0.34) | 0.33 (0.25 to 0.41) | −0.07 (−0.17 to 0.04) |
| Cervical cancer | 0.23 (0.16 to 0.30) | 0.20 (0.14 to 0.27) | 0.03 (−0.07 to 0.12) |
| Colorectal cancer | 0.20 (0.14 to 0.27) | 0.21 (0.14 to 0.28) | −0.01 (−0.10 to 0.08) |
| **Diabetes** | | | |
| Hemoglobin $A_{1c}$ test done | 0.77 (0.69 to 0.83) | 0.75 (0.68 to 0.82) | 0.01 (−0.08 to 0.11) |
| Hemoglobin $A_{1c}$ level <7% | 0.43 (0.35 to 0.51) | 0.43 (0.35 to 0.52) | −0.01 (−0.12 to 0.11) |
| Hemoglobin $A_{1c}$ level >9% or no test | 0.35 (0.27 to 0.43) | 0.35 (0.28 to 0.44) | −0.01 (−0.11 to 0.10) |
| LDL cholesterol level <2.59 mmol/L (<100 mg/dL) | 0.57 (0.49 to 0.65) | 0.37 (0.30 to 0.46) | 0.20 (0.09 to 0.31) |
| **Influenza vaccine, age ≥50 y** | 0.35 (0.28 to 0.44) | 0.30 (0.23 to 0.38) | 0.05 (−0.05 to 0.16) |
| **IVD** | | | |
| Appropriate antithrombotic medication | 0.75 (0.68 to 0.82) | 0.65 (0.57 to 0.73) | 0.10 (0.00 to 0.20) |
| LDL cholesterol level <2.59 mmol/L (<100 mg/dL) | 0.53 (0.44 to 0.61) | 0.43 (0.35 to 0.52) | 0.09 (−0.02 to 0.21) |
| **Pneumoccocal vaccination** | 0.27 (0.20 to 0.34) | 0.48 (0.40 to 0.56) | −0.21 (−0.32 to −0.11) |

IVD = ischemic vascular disease; LDL = low-density lipoprotein.

for validation will be needed. Fourth, we did not collect information on the reasons for differences between manual review and electronic reporting; however, other studies have explored that in detail, noting that electronic reporting frequently fails to capture evidence of appropriate care that is not documented in structured fields (10). Finally, this study was conducted before financial incentives for meaningful use were in place. With financial incentives, providers may be more motivated to document care in structured fields, which could increase accuracy. However, such incentives could also decrease accuracy if they increase attempts to game the system (23), which has been cited among potential unintended consequences of pay-for-performance and suggests the need for ongoing monitoring of reporting accuracy.

Despite these limitations, our study has important implications. It suggests that physicians need to recognize EHRs not as electronic versions of paper records but as tools that enable transformation in the way care is delivered, documented, measured, and improved. This is consistent with other studies that found that dictating notes into an EHR is associated with lower quality of care than using structured fields (24). Changing clinical workflow to support documentation in structured fields has been found to have a substantial effect on the accuracy of electronic reporting, enabling corrections of rates of recommended care by as much as 15 percentage points (from 50% to 65%) through documentation changes alone (25). The federal Regional Extension Center program can help providers learn how to use EHRs effectively, including assisting in workflow redesign (26).

Meaningful use measures overall, including those studied in depth here, are based on measures originally designed for manual review or claims. Thus, they represent only the beginning of what can be reported about quality from EHRs (27, 28). Future quality measures will be even more complex, incorporating more data elements and more complex data elements (6).

The National Quality Forum is developing specifications for how meaningful use clinical quality measures should be reported electronically (29). However, it is not clear that the specifications will be validated by direct comparisons of electronic reporting and manual review. If validated, the results could potentially be used to refine and retest the electronic reporting specifications until the reporting meets a minimum threshold for accuracy or to fine tune the list of clinical quality measures for financial incentives, selecting only those measures that were shown to have high sensitivity and specificity. The National Quality Forum eMeasures are designed to standardize reporting across EHR vendor products, but consistency across products has not yet been verified. Future studies (or future steps in the EHR certification process) could test electronic reporting on simulated charts for which rates of appropriate care are predetermined.

In conclusion, we found substantial measure-to-measure variability in the accuracy of 11 electronically reported meaningful use clinical quality measures. Practicing physicians are already concerned about the ability of quality reports to accurately represent the care they provide (30). If electronic reports are not proven to be accurate, their ability to change physicians' behavior to achieve higher quality, the underlying goal, will be undermined. This study suggests that national programs that link financial incentives to quality reporting should require that EHR vendors demonstrate the accuracy of their automated reports.

From the Center for Healthcare Informatics and Policy, Weill Cornell Medical College, Institute for Family Health, and New York-Presbyterian Hospital, New York, and Taconic Independent Practice Association, Fishkill, New York.

Presented in part at the Annual Symposium of the American Medical Informatics Association, Washington, DC, 22–26 October 2011.

**Note:** The authors had full access to all of the data in the study and take responsibility for the integrity of the data and accuracy of the data analysis.

**Acknowledgment:** The authors thank Jonah Piascik for his assistance with data collection.

**Grant Support:** By the Agency for Healthcare Research and Quality (grant R18 HS 017067).

**Potential Conflicts of Interest:** Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M12-1178.

**Reproducible Research Statement:** *Study protocol and statistical code:* Available from Dr. Kern (e-mail, lmk2003@med.cornell.edu). *Data set:* Not available.

**Requests for Single Reprints:** Lisa M. Kern, MD, MPH, Department of Public Health, Weill Cornell Medical College, 425 East 61st Street, Suite 301, New York, NY; e-mail, lmk2003@med.cornell.edu.

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

## References

1. **Blumenthal D.** Launching HITECH. N Engl J Med. 2010;362:382-5. [PMID: 20042745]
2. **Blumenthal D, Tavenner M.** The "meaningful use" regulation for electronic health records. N Engl J Med. 2010;363:501-4. [PMID: 20647183]
3. **Steinbrook R.** Health care and the American Recovery and Reinvestment Act. N Engl J Med. 2009;360:1057-60. [PMID: 19224738]
4. **Centers for Medicare & Medicaid Services (CMS), HHS.** Medicare and Medicaid programs; electronic health record incentive program. Final rule. Fed Regist. 2010;75:44313-588. [PMID: 20677415]
5. **Centers for Medicare & Medicaid Services (CMS), HHS.** Medicare and Medicaid programs; electronic health record incentive program—stage 2. Final rule. Fed Regist. 2012;77:53967-4162. [PMID: 22946138]
6. **Bates DW.** The approaching revolution in quality measurement [Editorial]. Jt Comm J Qual Patient Saf. 2009;35:358. [PMID: 19634803]

7. **Baker DW, Persell SD, Thompson JA, Soman NS, Burgner KM, Liss D, et al.** Automated review of electronic health records to assess quality of care for outpatients with heart failure. Ann Intern Med. 2007;146:270-7. [PMID: 17310051]
8. **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]
9. **Kmetik KS, O'Toole MF, Bossley H, Brutico CA, Fischer G, Grund SL, et al.** Exceptions to outpatient quality measures for coronary artery disease in electronic health records. Ann Intern Med. 2011;154:227-34. [PMID: 21320938]
10. **Parsons A, McCullough C, Wang J, Shih S.** Validity of electronic health record-derived quality measurement for performance monitoring. J Am Med Inform Assoc. 2012;19:604-9. [PMID: 22249967]
11. **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]
12. **Weiner M, Stump TE, Callahan CM, Lewis JN, McDonald CJ.** Pursuing integration of performance measures into electronic medical records: beta-adrenergic receptor antagonist medications. Qual Saf Health Care. 2005;14:99-106. [PMID: 15805454]
13. **Kern LM, Dhopeshwarkar R, Barrón Y, Wilcox A, Pincus H, Kaushal R.** Measuring the effects of health information technology on quality of care: a novel set of proposed metrics for electronic quality reporting. Jt Comm J Qual Patient Saf. 2009;35:359-69. [PMID: 19634804]
14. The Institute for Family Health. Accessed at www.institute2000.org on 9 November 2012.
15. **Centers for Medicare & Medicaid Services.** Quality Measures. Baltimore, MD: Centers for Medicare & Medicaid Services; 2012. Accessed at www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures on 12 November 2012.
16. **Gordis L.** Assessing the validity and reliability of diagnostic and screening tests. In: Gordis L. Epidemiology, 4th ed. Philadelphia: Saunders Elsevier; 2008: 85-108.
17. **Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al.** Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA. 2011;306:848-55. [PMID: 21862746]
18. **Nadkarni PM, Ohno-Machado L, Chapman WW.** Natural language processing: an introduction. J Am Med Inform Assoc. 2011;18:544-51. [PMID: 21846786]
19. **Chan KS, Fowles JB, Weiner JP.** Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev. 2010;67:503-27. [PMID: 20150441]
20. **Linder JA, Kaleba EO, Kmetik KS.** Using electronic health records to measure physician performance for acute conditions in primary care: empirical evaluation of the community-acquired pneumonia clinical quality measure set. Med Care. 2009;47:208-16. [PMID: 19169122]
21. **Roth CP, Lim YW, Pevnick JM, Asch SM, McGlynn EA.** The challenge of measuring quality of care from the electronic health record. Am J Med Qual. 2009;24:385-94. [PMID: 19482968]
22. **Tang PC, Ralston M, Arrigotti MF, Qureshi L, Graham J.** Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. J Am Med Inform Assoc. 2007;14:10-5. [PMID: 17068349]
23. **Mannion R, Davies HT.** Payment for performance in health care. BMJ. 2008;336:306-8. [PMID: 18258966]
24. **Linder JA, Schnipper JL, Middleton B.** Method of electronic health record documentation and quality of primary care. J Am Med Inform Assoc. 2012;19:1019-24. [PMID: 22610494]
25. **Baron RJ.** Quality improvement with an electronic health record: achievable, but not automatic. Ann Intern Med. 2007;147:549-52. [PMID: 17938393]
26. Office of the National Coordinator for Health Information Technology. Regional Extension Centers. Washington, DC: U.S. Department of Health and Human Services; 2012. Accessed at www.healthit.gov/providers-professionals/regional-extension-centers-recs on 12 November 2012.
27. **Pawlson LG.** The past as prologue: future directions in clinical performance measurement in ambulatory care. Am J Manag Care. 2007;13:594-6. [PMID: 17988184]
28. **Weiner JP, Fowles JB, Chan KS.** New paradigms for measuring clinical performance using electronic health records. Int J Qual Health Care. 2012;24:200-5. [PMID: 22490301]
29. National Quality Forum Electronic Quality Measures (eMeasures). Washington, DC: National Quality Forum; 2012. Accessed at www.qualityforum.org/Projects/e-g/eMeasures/Electronic_Quality_Measures.aspx on 12 November 2012.
30. **Ofri D.** Quality measures and the individual physician. N Engl J Med. 2010;363:606-7. [PMID: 20818853]

**Current Author Addresses:** Drs. Kern and Kaushal and Ms. Edwards: Center for Healthcare Informatics and Policy, Weill Cornell Medical College, 425 East 61st Street, Suite 301, New York, NY 10065.
Dr. Malhotra: Weill Cornell Medical College, 575 Lexington Avenue, Box 110, New York, NY 10022.
Ms. Barrón: Center for Home Care and Research, Visiting Nurse Service of New York, 1250 Broadway, 20th Floor, New York, NY 10001.
Ms. Quaresimo: 4 Cleveland Drive, Poughkeepsie, NY 12601.
Ms. Dhopeshwarkar: 2665 Prosperity Avenue, Apartment 337, Fairfax, VA 22031.
Ms. Pichardo: Institute for Family Health, 22 West 19th Street, 8th Floor, New York, NY 10011.

**Author Contributions:** Conception and design: L.M. Kern, S. Malhotra, R. Kaushal.
Analysis and interpretation of the data: L.M. Kern, S. Malhotra, Y. Barrón, R. Dhopeshwarkar, M. Pichardo, A.M. Edwards, R. Kaushal.
Drafting of the article: L.M. Kern, S. Malhotra, M. Pichardo, R. Kaushal.
Critical revision of the article for important intellectual content: L.M. Kern, S. Malhotra, Y. Barrón, R. Dhopeshwarkar, R. Kaushal.
Final approval of the article: L.M. Kern, Y. Barrón, A.M. Edwards, R. Kaushal.
Provision of study materials or patients:
Statistical expertise: Y. Barrón, A.M. Edwards.
Obtaining of funding: L.M. Kern, R. Kaushal.
Administrative, technical, or logistic support: S. Malhotra, R. Dhopeshwarkar, M. Pichardo.
Collection and assembly of data: L.M. Kern, S. Malhotra, Y. Barrón, J. Quaresimo, M. Pichardo.

## *Appendix Table.* Measure Specifications

| Metric | Numerator | Denominator |
|---|---|---|
| 1. Asthmatic population: asthma control | Patients who were prescribed either the preferred long-term control medication (inhaled corticosteroid) or an acceptable alternative treatment (leukotriene modifiers, cromolyn sodium, nedocromil sodium, or sustained-release methylxanthines) (a link to a drug list by the AMA was provided). | All patients aged 5–40 y with mild, moderate, or severe persistent asthma. Patient selection: ICD-9-CM codes for asthma 493.00–493.92; additional individual medical record review must be completed to identify those patients.<br>Exclusions: Documentation of reason(s) for not prescribing either the preferred long-term control medication (inhaled corticosteroid) or an acceptable alternative treatment. |
| 2. Preventive care population: breast cancer screening | Number of patients in denominator who had a mammogram (ordered or self-reported) within 24 mo up to and including the last day of the reporting period. NCQA measure uses the following numerator codes: CPT codes 76083, 76090–76092; ICD-9-CM codes 87.36, 87.37; V codes V76.11, V76.12; UB-92 codes 0401, 0403).<br>Documentation in the medical record must include both a note indicating the date the mammogram was done and the result or finding. | Number of unique female patients aged 52–69 y having at least 1 visit in the previous 12 mo up to and including the last day of the reporting period.<br>Exclude women who had a bilateral mastectomy. If there is evidence of 2 separate mastectomies, this patient may be excluded from the measure. The bilateral mastectomy must have occurred by the last day of the measurement year. (For bilateral: ICD-9-CM codes 85.42, 85.44, 85.46, 85.48; CPT codes 19180.50 or 19180 with modifier 09950*, 19200.50 or 19200 with modifier code 09950*, 19220.50 or 19220 with modifier 09950*, 19240.50 or 19240 with modifier 09950*. For unilateral [need 2 separate occurrences on 2 different dates of service]: ICD-9-CM codes 85.41, 85.43, 85.45, 85.47; CPT codes 19180, 19200, 19220, 19240.) |
| 3. Preventive care population: cervical cancer screening | Number of patients in denominator having had a cervical cancer screening test (Pap test) within 36 mo up to and including the last day of the reporting period.<br>Documentation in the medical record must include a note indicating the date the test was done and the result or finding. | Number of unique female patients aged 21–64 y having at least 1 visit in the previous 12 mo up to and including the last day of the reporting period.<br>Exclude women who had a hysterectomy and have no residual cervix. Exclusionary evidence in the medical record must include a note indicating a hysterectomy with no residual cervix. Documentation of "complete hysterectomy," "total hysterectomy," "total abdominal hysterectomy," or "radical hysterectomy" meets the criteria. Documentation of "hysterectomy" alone does not meet the criteria because it does not indicate that the cervix has been removed. The hysterectomy must have occurred by the last day of the measurement year. Use any of the following codes or descriptions of codes in the medical record to identify allowable exclusions: Surgical codes for hysterectomy (CPT codes 51925, 56308, 58150, 58152, 58200, 58210, 58240, 58260, 58262, 58263, 58267, 58270, 58275, 58280, 58285, 58290–58294, 58550, 58551, 58552–58554, 58951, 58953–58954, 58956, 59135; ICD-9-CM codes 68.4–68.8, 618.5; V codes V67.01, V76.47). |
| 4. Preventive care population: colorectal cancer screening | Number of patients in denominator having 1 or more of the following documented completed tests: a fecal occult blood test within 12 mo up to and including the last day of the reporting period, a flexible sigmoidoscopy within 5 y up to and including the last day of the reporting period, a double contrast barium enema within 5 y up to and including the last day of the reporting period, or a colonoscopy within 10 y up to and including the last day of the reporting. Documentation in the medical record must include both a note indicating the date on which the colorectal cancer screening was done and, for a notation in the progress notes, the result or finding (this ensures that the screening was done and not merely ordered).<br>For a notation in the medical history, a result is not required. Documentation in the medical history pertains to screenings that happened in the past and it is assumed that the result was negative (a positive result would have been noted as such). A notation in the medical history must include a date reference that meets the timeline outlined in the specifications. | Number of unique patients aged 51–80 y with at least 1 visit in past 12 mo |
| 5. Diabetes population: hemoglobin $A_{1c}$ testing | Number of patients in denominator who had 1 or more hemoglobin $A_{1c}$ test results recorded during the past 12 mo up to and including the last day of the reporting period (can be identified by either CPT code 83036 or LOINC codes 4548-4, 4549-2, 17855-8, 17856-6, or 4637-5, or an automated laboratory record with a service date, or, at minimum, documentation in the medical record must include a note indicating the date on which the hemoglobin $A_{1c}$ test was done and the result). | Number of unique patients seen in the reporting period, aged 18–75 y, with 2 ambulatory care visits since diabetes diagnosis in past 24 mo. Diabetes diagnosis: ICD-9-CM codes 250, 357.2, 362.0, 366.41, 648.0; DRGs 294, 205. Outpatient or nonacute inpatient: CPT codes 92002–92014, 99201–99205, 99211–99215, 99217–99220, 99241, 99245, 99271–99275, 99301–99303, 99311–99313, 99321–99323, 99331–99333, 99341–99355, 99384–99387, 99394–99397, 99401–99404, 99411, 99412, 99420, 99429, 99499; UB-92 revenue codes 019X, 0456, 049X–053X, 055X–059X, 065X, 066X, 076X, 077X, 082X–085X, 088X, 092X, 094X, 096X, 0972–0979, 0982–0986, 0988, 0989. |

*Continued on following page*

*Appendix Table*—**Continued**

| Metric | Numerator | Denominator |
|---|---|---|
| 6. Diabetes population: hemoglobin $A_{1c}$ levels (good control) | Number of patients in denominator having at least 1 hemoglobin $A_{1c}$ level measured in the past 12 mo up to and including the last day of the reporting period and whose most recent recorded hemoglobin $A_{1c}$ level is <7%. | Number of unique patients seen in the reporting period, aged 18–75 y, with 2 ambulatory care visits since diabetes diagnosis in past 24 mo. Diabetes diagnosis: ICD-9-CM codes 250, 357.2, 362.0, 366.41, 648.0; DRGs 294, 205. Outpatient or nonacute inpatient: CPT codes 92002–92014, 99201–99205, 99211–99215, 99217–99220, 99241, 99245, 99271–99275, 99301–99303, 99311–99313, 99321–99323, 99331–99333, 99341–99355, 99384–99387, 99394–99397, 99401–99404, 99411, 99412, 99420, 99429, 99499; UB-92 revenue codes 019X, 0456, 049X–053X, 055X–059X, 065X, 066X, 076X, 077X, 082X–085X, 088X, 092X, 094X, 096X, 0972–0979, 0982–0986, 0988, 0989. |
| 7. Diabetes population: hemoglobin $A_{1c}$ levels (poor control) | Number of patients in denominator having at least 1 hemoglobin $A_{1c}$ level measured in the past 12 mo up to and including the last day of the reporting period and whose most recent recorded hemoglobin $A_{1c}$ level is >9%, plus the number of patients in denominator who have had no hemoglobin $A_{1c}$ levels measured in the previous 12 mo up to and including the last day of the reporting period. | Number of unique patients seen in the reporting period, aged 18–75 y, with 2 ambulatory care visits since diabetes diagnosis in past 24 mo. Diabetes diagnosis: ICD-9-CM codes 250, 357.2, 362.0, 366.41, 648.0; DRGs 294, 205. Outpatient or nonacute inpatient: CPT codes 92002–92014, 99201–99205, 99211–99215, 99217–99220, 99241, 99245, 99271–99275, 99301–99303, 99311–99313, 99321–99323, 99331–99333, 99341–99355, 99384–99387, 99394–99397, 99401–99404, 99411, 99412, 99420, 99429, 99499; UB-92 revenue codes 019X, 0456, 049X–053X, 055X–059X, 065X, 066X, 076X, 077X, 082X–085X, 088X, 092X, 094X, 096X, 0972–0979, 0982–0986, 0988, 0989. |
| 8. Diabetes population: LDL cholesterol levels <2.59 mmol/L (<100 mg/L) | Number of patients in denominator having at least 1 LDL cholesterol level measured in the past 12 mo up to and including the last day of the reporting period and whose most recent recorded LDL cholesterol level is <2.59 mmol/L (<100 mg/L). | Number of unique patients seen in the reporting period, aged 18–75 y, with 2 ambulatory care visits since diabetes diagnosis in past 24 mo. Diabetes diagnosis: ICD-9-CM codes 250, 357.2, 362.0, 366.41, 648.0; DRGs 294, 205. Outpatient or nonacute inpatient: CPT codes 92002–92014, 99201–99205, 99211–99215, 99217–99220, 99241, 99245, 99271–99275, 99301–99303, 99311–99313, 99321–99323, 99331–99333, 99341–99355, 99384–99387, 99394–99397, 99401–99404, 99411, 99412, 99420, 99429, 99499; UB-92 revenue codes 019X, 0456, 049X–053X, 055X–059X, 065X, 066X, 076X, 077X, 082X–085X, 088X, 092X, 094X, 096X, 0972–0979, 0982–0986, 0988, 0989. |
| 9A. Preventive care population: flu shots, aged 50–64 y | Number of patients in denominator who received a flu shot since the most recent 1 September. | Number of unique patients aged 50–64 y seen for at least 1 visit in the previous 24 mo up to and including the last day of the reporting period. |
| 9B. Preventive care population: flu shots, aged >64 y | Number of patients in denominator who received a flu shot since the most recent 1 September. | Number of unique patients at aged ≥65 y seen for at least 1 visit in the previous 24 mo up to and including the last day of the reporting period. |
| 10. Cardiovascular disease population: use of aspirin or another antithrombotic in patients with IVD | Number of patients who have documentation of use of aspirin or another antithrombotic during the 12-mo measurement period. Documentation in the medical record must include, at a minimum, a note indicating the date on which aspirin or another antithrombotic was prescribed or documentation of prescription from another treating physician. (Exclude patient self-report.) | Number of patients aged ≥18 y with a diagnosis of IVD who have been under the care of the physician or physician group for IVD for at least 12 mo (this is defined by documentation of a face-to-face visit for IVD care between the physician and patient that predates the most recent IVD visit by at least 12 mo). Codes to identify a patient with a diagnosis of IVD: ICD-9-CM codes 411, 413, 414.0, 414.8, 414.9, 429.2, 433–434, 440.1, 440.2, 444, 445. |
| 11. Cardiovascular disease population: LDL cholesterol levels <2.59 mmol/L (<100 mg/dL) in patients with IVD | Number of patients in denominator having at least 1 LDL cholesterol level measured in the past 12 mo up to and including the last day of the reporting period and a recorded LDL cholesterol level <2.59 mmol/L (<100 mg/L) in the past 12 mo. (Exclude patient self-report or self-monitoring, LDL–HDL ratio, and findings reported on progress notes or other nonlaboratory documentation). | Number of patients aged ≥18 y with a diagnosis of IVD who have been under the care of the physician or physician group for IVD for at least 12 mo (this is defined by documentation of a face-to-face visit for IVD care between the physician and patient that predates the most recent IVD visit by at least 12 mo). Codes to identify a patient with a diagnosis of IVD: ICD-9-CM codes 411, 413, 414.0, 414.8, 414.9, 429.2, 433–434, 440.1, 440.2, 444, 445. |
| 12. Preventive care population: pneumococcal vaccination | Number of patients in denominator who have ever received the pneumococcal vaccination (CPT code 90732). | Number of unique patients aged ≥65 y seen for at least 1 visit in the reporting period.<br>Exclusions: previous anaphylactic reaction to the vaccine or components; other medical reason(s) documented by the practitioner for not receiving a pneumococcal vaccination (ICD-9-CM exclusion codes for PC-8 pneumonia vaccination: 995.0 and E949.6, 995.1 and E949.6, 995.2 and E949.6); and patient reason(s) (e.g., economic, social, religious). |

AMA = American Medical Association; CPT = Current Procedural Terminology; DRG = diagnosis-related group; HDL = high-density lipoprotein; ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification; IVD = ischemic vascular disease; LDL = low-density lipoprotein; LOINC = Logical Observation Identifiers Names and Codes; NCQA = National Committee for Quality Assurance; UB = Uniform Billing.
* .50 and 09950 modifier codes indicate that the procedure was bilateral and done during the same operative session.