Using Computer-Extracted Data from Electronic Health Records to Measure the Quality of Adolescent Well-Care

William Gardner, Suzanne Morton, Sepheen C. Byron, Aldo Tinoco, Benjamin D. Canan, Karen Leonhart, Vivian Kong, and Sarah Hudson Scholle

Objective. To determine whether quality measures based on computer-extracted EHR data can reproduce findings based on data manually extracted by reviewers.

Data Sources. We studied 12 measures of care indicated for adolescent well-care visits for 597 patients in three pediatric health systems.

Study Design. Observational study.

Data Collection/Extraction Methods. Manual reviewers collected quality data from the EHR. Site personnel programmed their EHR systems to extract the same data from structured fields in the EHR according to national health IT standards.

Principal Findings. Overall performance measured via computer-extracted data was 21.9 percent, compared with 53.2 percent for manual data. Agreement measures were high for immunizations. Otherwise, agreement between computer extraction and manual review was modest ($Kappa = 0.36$) because computer-extracted data frequently missed care events (sensitivity = 39.5 percent). Measure validity varied by health care domain and setting. A limitation of our findings is that we studied only three domains and three sites.

Conclusions. The accuracy of computer-extracted EHR quality reporting depends on the use of structured data fields, with the highest agreement found for measures and in the setting that had the greatest concentration of structured fields. We need to improve documentation of care, data extraction, and adaptation of EHR systems to practice workflow.

Key Words. Quality measurement, pediatric well-care, electronic health records

The United States is making an extensive investment in electronic health records (EHRs) with the goal of improving quality of care and our ability to monitor it (U.S. Department of Health and Human Services 2011). EHRs also have enormous potential for facilitating practice-based research, but the feasibility of such use depends on the validity of the quality data in the EHR

(Wasserman 2011). “Feasibility” is the extent to which the required data can be extracted for quality measurement without undue burden. Quality measurement now relies largely on administrative claims data, but measures for evaluating the clinical content of preventive care have not been widely implemented because of the need for more detailed clinical data than are readily available in administrative claims (Iezzoni 1997; MacLean et al. 2006). Calculating quality measures using data extracted from EHRs provides more clinical details and better quality measurements than quality measures based on administrative data (Tang et al. 2007). Feasible EHR-based quality measures, however, must use data extracted from the EHR by computer rather than by costly manual extraction (Roth et al. 2009).

Before quality measures calculated from computer-extracted data can be used for judging performance, these measures must be shown to be valid. Specifically, we are examining the concurrent validity of automatically extracting EHR data for use in the measurement of the quality of adolescent well-care. Concurrent validity means that we expect that information that a computer has extracted from an EHR would be similar to information that a human has recorded from the same EHR.

Previous efforts to validate computer-extracted EHR quality measurement have a mixed record and do not resolve the questions around potential factors for success or failure of prior validation efforts (Weiner et al. 2005; Persell et al. 2006; Fuhlbrigge et al. 2008; Linder, Kaleba, and Kmetik 2009; Gold et al. 2012; Hazelhurst et al. 2012). Why should the validity of quality measures calculated from computer-extracted data vary across studies? One reason is that the feasibility of using computers to extract quality data may vary across the sites at which the studies were carried out and the different clinical domains that were being measured. Data recorded in structured fields are more readily extracted from an EHR than data recorded in free text notes. Hence, EHR system developers can increase the feasibility of quality data extraction by providing structured fields for recording clinical data or by making it easy for sites to build such fields into the EHR to accommodate local provider workflow.

Address correspondence to William Gardner, Ph.D., Department of Pediatrics, Dalhousie University, 5850-5980 University Ave, Halifax, NS B3K 6R8; e-mail: william.p.gardner@gmail.com. Suzanne Morton, M.P.H., M.B.A., Sepheen C. Byron, M.H.S., Aldo Tinoco, M.D., Ph.D., Vivian Kong, M.P.H., and Sarah Hudson Scholle, M.P.H., Dr.P.H., are with the National Committee for Quality Assurance, Washington, DC. Benjamin D. Canan, M.P.H., and Karen Leonhart, B.S., are with the Center for Innovation in Pediatric Practice, The Research Institute at Nationwide Children’s Hospital, Columbus, OH.

The Office of the National Coordinator for Health Information Technology of the US Department of Health and Human Services has established objectives and standards for the use of health information technology in the assessment of health care, along with EHR criteria needed to achieve these objectives. According to these criteria, the data used for quality measurement must be available in structured fields using standardized health terminologies (U.S. Department of Health and Human Services 2012).

In this study, we measured the agreement of quality measures calculated from data extracted by computer from EHRs (hereafter, “computer-extracted measures”) by comparing them against data obtained by manual reviewers examining the same EHR (hereafter, “manually extracted measures”). Data were collected at three sites that varied in the degree to which quality data were captured in structured fields. We expected that the agreement of measures calculated from computer-extracted data would be higher for quality measures that involve data elements represented in the EHR by structured fields. Similarly, we expected that agreement would be higher at sites where structured fields were more prevalent. We used measures that assess health care given to adolescents in three health care domains: (a) immunizations, (b) screenings, and (c) behavioral health risk assessment and counseling. Before collecting data on the performance of services, we asked personnel at the sites about how their EHRs functioned, including whether the data elements required for the quality measures were implemented as structured database fields. In summary, this study assessed the concurrent validity of measures of the quality of preventive care for adolescents and examined variation in such validity as a function of the site and the quality domain being measured.

METHODS

Design

This study assessed the concurrent validity of quality measures based on computerized extracts of EHRs (computer-extracted measures) by comparing them with data extracted by human reviews of the same records (manually extracted measures).

Settings

Three U.S. pediatric health care organizations were selected for study inclusion based on the availability of data within EHRs. Institutions provided

primary care services and had implemented an EHR system at least 2 years prior to our study. In addition, the sites were selected to represent several practice types (family practice, general pediatrics, and adolescent medicine), practice settings (children's hospitals, private practices, and clinics serving vulnerable youth), EHR systems, and patient populations. Each site used a different EHR vendor. Site 1 was a network of clinics serving homeless and vulnerable adolescents and used the eClinicalWorks system. Patients were sampled from three clinics: one clinic from a major Eastern city and two from major Southwestern cities. Site 2 was a large Midwestern pediatric hospital and used the Epic system. The patients were sampled from a Medicaid managed care organization associated with the hospital. Site 3 was an adolescent medicine clinic primarily providing behavioral health and gynecology care to young women and was affiliated with a Midwestern children's hospital. This clinic used the AllScripts system.

Patients

Adolescents whose age was ≥ 12 years and ≤ 19 years on December 31, 2010 qualified for the study. However, because site 2 served adolescents through age 18, this sample was limited to adolescents between 12 and 18 years of age. Each adolescent must have had a face-to-face visit to a primary care practice at the organization between January 1, 2011 and December 31, 2011 and a face-to-face visit to the same practice between January 1, 2010 and December 31, 2010. Each site was asked to randomly select 200 patients who met these criteria. Only one site, however, was able to identify more than 200 eligible patients. This site used the randomization function in Excel to choose the sample. A second site could only provide 197 adolescents meeting the criteria, and the third site similarly reported that they could identify only 200 patients. Therefore, neither of the latter sites was able to randomly select patients. The total $n = 597$; Table 1 describes the patient demographics.

Measures of the Quality of Adolescent Well-Care

We tested 12 measures of the quality of adolescent well-care. Measures were derived from quality indicators developed by the National Committee for Quality Assurance and modified as required for this study, as described in Table 2. These measures covered diverse aspects of adolescent preventive care and fell into three domains: *immunizations* (three measures), *screening*

Table 1: Patient Demographics by Site

<i>Characteristic</i>	<i>Site 1</i>	<i>Site 2</i>	<i>Site 3</i>	<i>Total</i>
<i>N</i>	200	200	197	597
Mean age (range), years	16.9 (13.0, 20.5)	15.8 (13.0, 18.9)	17.7 (14.1, 20.4)	16.8 (13.0, 20.5)
Females (%)	47.0	59.5	98.5	68.2
Insurance (%)				
Medicaid	60.5	100.0	42.1	34.2
Children's Health Insurance Program	0.0	0.0	0.0	0.0
Commercial	0.5	0.0	50.3	16.8
Self-pay	30.0	0.0	1.5	10.6
Other	7.5	0.0	0.5	2.7
Not documented	1.5	0.0	5.6	2.3
Race/ethnicity (%)				
White, Non-Hispanic	10.0	28.0	52.3	30.0
African-American, Non-Hispanic	35.0	57.5	40.6	44.4
Latino/Hispanic	19.5	1.0	0.5	7.0
Asian, Native American, or Pacific Islander	3.0	1.0	0.0	1.3
Other	27.5	9.5	1.5	12.9
Not documented	5.0	3.0	5.1	4.4
English spoken at home (%)	90.5	92.5	95.4	92.8
Mean number of special health conditions (range)	0.5 (0, 4)	0.5 (0, 6)	0.6 (0, 4)	0.4 (0, 7)

(three measures), and *assessment and counseling for risky behaviors* (six measures). A detailed description of each measure is available as an online appendix.

Data Collection Methods

Manually Extracted Data. Trained reviewers from the research team conducted manual EHR reviews at each site. They collected information on the quality measures, visit history, and sociodemographic characteristics of the patients. The review focused on care that occurred from October 1, 2010 to December 31, 2011 (a 15-month observation period).

Two reviewers independently collected data for the same 75 adolescent records across the three sites to assess inter-rater reliability. They coded 56 data elements from the 12 measures. Inter-rater reliability coefficients could not be calculated for 12 (21.4 percent) of these elements because there were no service events detected. For the other elements, the manual reviewers were

Table 2: Definitions of Quality Measures

Measure	Denominator	Numerator	Reference
Immunizations domain			
HPV vaccination	Females aged 13–14 years	Did the adolescent receive all three doses of the HPV vaccination by age 13?	NCQA HEDIS measure <i>HPV Vaccination for Female Adolescents*</i>
Meningococcal vaccination	Adolescents aged 13–14 years	Did the adolescent receive the Meningococcal vaccine by age 13?	NCQA HEDIS measure <i>Immunizations for Adolescents†</i>
Tetanus, diphtheria, and pertussis (Tdap/Td) vaccination	Adolescents aged 13–14 years	Did the adolescent receive the Tdap/Td vaccine by age 13?	NCQA HEDIS measure <i>Immunizations for Adolescents†</i>
Screenings domain			
Blood pressure screening	Adolescents aged 12–20 years	Was a blood pressure screening performed and results documented?	NCQA measure <i>Blood Pressure Screening by Age 18 Years, by Age 13 Years</i>
Chlamydia screening	Females aged 12–20 years who are sexually active	Was chlamydia screening performed?	NCQA HEDIS measure <i>Chlamydia Screening in Women‡</i>
Depression screening	Adolescents aged 12–20 years, except those already diagnosed with depression	Was depression screening using a standardized tool performed?	NCQA measure <i>Depression Screening by Age 18 Years</i>

continued

Table 2. Continued

Measure	Denominator	Numerator	Reference
Risky behavior assessment/counseling domain Nutrition counseling	Adolescents aged 12–17 years	Has the adolescent been counseled about nutrition?	NCQA HEDIS measure <i>Weight Assessment and Counseling for Nutrition and Physical Activity for Children/Adolescents</i> [§]
	Adolescents aged 12–17 years	Has the adolescent been counseled about being physically active?	NCQA HEDIS measure <i>Weight Assessment and Counseling for Nutrition and Physical Activity for Children/Adolescents</i> [§]
Physical activity counseling	Adolescents aged 12–17 years		NCQA-developed measure under NCINQ ₂
Sexual activity documentation	Adolescents aged 12–20 years	Is there an indication about whether the adolescent is sexually active?	NCQA measure <i>Risky Behavior Assessment/Counseling by Age 18 Years</i> [‡]
Alcohol use assessment	All adolescents aged 12–20 years	Is there an indication about whether the adolescent is drinking?	NCQA measure <i>Risky Behavior Assessment/Counseling by Age 18 Years</i> [‡]
Substance use assessment	All adolescents aged 12–20 years	Is there an indication about whether the adolescent is using drugs?	NCQA measure <i>Risky Behavior Assessment/Counseling by Age 18 Years</i> [‡]

continued

Table 2. Continued

Measure	Denominator	Numerator	Reference
Tobacco use and help with quitting	Adolescents aged 12–20 years	Is there an indication about whether the adolescent is a tobacco user, and if the adolescent is a user, did he/she receive advice about quitting, counseling, or referral/enrollment in a cessation program?	NCQA-developed measure under NCINQ

Note. In some cases we modified a measure developed for another context.

- *The original measure denominator included female adolescents age 13. The denominator used for this study was female adolescents ages 13 and 14.
- †The original measure denominator included adolescents age 13. The denominator used for this study was adolescents ages 13 and 14.
- *The original measure assesses adolescent females ages 16–20 with an indication for sexual activity as determined by claims data. The denominator used for this study was females ages 12–20. We also used current/past sexual activity status, number of sexual partners, and use of nonpharmaceutical methods of contraception as indicators of sexual activity.
- §The original measure denominator included children and adolescents age 3–17 years of age. The denominator used for this study was adolescents age 12–17 years only.
- †The original measure assesses the percentage of adolescents with documentation of a risk assessment or counseling for risky behaviors by the age of 18 years. The measure includes four risky behavior topics assessed as four rates: Alcohol Use, Tobacco Use, Substance Use, and Sexual Activity, and the numerator was expressed as assessment or counseling. The numerator used in this study was for assessment only.

highly reliable: the average Kappa for inter-rater agreement was 0.90 (SD = 0.18).

Computer-Extracted Data. Information services personnel at each site programmed database queries against structured data fields for the same patients and data elements abstracted by manual review. We provided a detailed data layout and instructions for the computerized extract from the EHR and trained personnel at the field site on the data collection and submission procedures. We provided a list of example values and codes that fulfilled each data element, including potential locations of this information in the EHR. Site personnel supplemented this list based on their local data dictionaries. We did not provide further specification of how the extraction was to be programmed. Neither did we inspect the extraction code. The site personnel had no access to the manually collected data. The research team linked the manual data and the computer-extracted data using common study identifiers.

Use of Structured Fields in the EHR. We also asked the site personnel to identify which of the required data elements for the quality measures were available in structured fields. This allowed us to distinguish whether the lack of agreement between manual review and computer extracts is due to a technical issue (lack of structured field) or workflow (the structured field is not used). For each quality measure at each site, we measured the extent that structured data was used by dividing the number of data elements represented as structured fields by the number of data elements in the measure specification.

Ethical Review

The Institutional Review Board at The Research Institute at Nationwide Children's Hospital reviewed this research. The Board judged that the research involved not greater than minimal risk and waived the requirement for informed consent.

Statistical Analysis

We evaluated whether performance results based on computer-extracted data could accurately reproduce results based on manual review of the EHR in four steps. First, we used Kappa statistics to determine whether, for each

measure and site, the manual review and computer extracts identified the same patients as meeting the denominator criteria of each measure. These analyses use the manually extracted data as a “gold standard” for both the denominator and numerator of the measure (Weiner et al. 2005). While this is not a true gold standard because the manual review itself can be a highly imperfect record of the care actually delivered, manually extracted data often are used as comparison data in validation studies (Persell et al. 2006; Baker et al. 2007; Parsons et al. 2012).

The second step was to test the concurrent validity of the quality measures calculated from computer-extracted data. To do this, we calculated the agreement between manual and computer-extracted data using Cohen’s Kappa. Kappa = 1.0 when the computer-extracted and manual quality data agree perfectly on an individual patient by individual patient basis and Kappa = 0.0 when agreement is at chance level. We also examined whether Kappa varied as a function of either the site where the patient received care or the measure domain (i.e., immunizations, screenings, or risky behavior assessments/counseling).

Third, to identify the specific kinds of patient-level disagreements that produce low Kappa (that is, low agreement), we calculated the sensitivity and specificity for the performance results based on computer extracts with respect to “gold standard” results based on manual review. Thus, the true positives were the adolescents who received recommended care according to the manual review, and the sensitivity is the proportion of those true positive patients who were identified as receiving the care in the computer extract. Similarly, the true negatives were the adolescents who did not receive recommended care according to the manual review, and the specificity is the proportion of those true negative patients who were identified as not receiving the recommended care in the computer extract.

Finally, we examined how the lack of agreement between computer-extracted data and manual data induced bias in the performance rates on the quality measures. The performance rate is the proportion of patients who should have received care (the denominator) who were reported as receiving care (the numerator). We measured overall bias and examined how the bias in performance rates based on computer-extracted data varied as a function of the site and the measure domain. To measure the bias in the performance rate calculated from computer-extracted data versus the performance rate calculated from manual review data, we needed a normalized metric. Therefore, we calculated

$$\text{Relative Performance} = 100\% \times \frac{\text{Performance based on Computer Extracts}}{\text{Performance based on Manual Review}}$$

To compare two sites (or two measurement domains), we calculated the difference between the Relative Performance rates for each site (or measurement domain).

In this study, each patient contributed a cluster of 12 observations (one for each quality measure). To estimate confidence intervals around reported statistics that accounted for this clustering, we carried out bootstrapping analyses using 10,000 replications.

RESULTS

Structured Fields

Table 3 presents information on whether data elements at the sites are captured in structured fields. Sites varied in the rates of implementation of structured fields, with site 1 having significantly more structured fields than site 3 ($p < .001$). Immunization measures were also more likely to have structured fields than the screening measures, and screening measures were more likely than the risky behavior assessment/counseling measures, but neither of these differences was statistically significant.

Agreement on Measure Denominators

There was relatively little disagreement between the manual reviews and the computer extracts for measure denominators. Overall, there were 12 quality measures on 597 patients. This means that there were $12 \times 597 = 7,164$ potential quality observations that could contribute data to the denominator. In the manual review data, adolescents actually qualified for the denominator for 68.4 percent of the opportunities, which was nearly identical to the 68.3 percent for the computer-extracted data. Agreement between the manual and computer-extracted denominator data was very high: Kappa = 0.95, 95 percent CI = [0.94, 0.96]. Looked at measure by measure, the agreement between the manual and computer-extracted measure denominators was greater than or equal to Kappa = 0.99 for all but two measures. The Kappa statistic for the chlamydia denominator was 0.78 (95 percent CI = [0.73, 0.83]). The chlamydia measure applies to sexually active girls; manual coding

Table 3: Percentages of Required Data Elements Found in Structured Fields for Each Measure, by Measure and Site

<i>Measure</i>	<i>Percentages of Required Data Elements Found in Structured Fields</i>				
	<i>Data Elements Required for Measure*</i>	<i>Site 1</i>	<i>Site 2</i>	<i>Site 3</i>	<i>Domain Average</i>
Immunizations domain					
HPV vaccination	3	100	100	100	78
Meningococcal vaccination	1	100	100	0	
Tetanus, diphtheria, and pertussis (Tdap/Td) vaccination	1	100	100	0	
Screening domain					
Blood pressure screening	2	50	100	50	72
Chlamydia screening	12	83	92	42	
Depression screening	3	100	33	100	
Counseling and risky behavior assessments domain					
Nutrition counseling	1	100	0	100	65
Physical activity counseling	1	0	0	100	
Sexual activity documentation	9	89	89	22	
Alcohol use assessment	1	100	100	100	
Substance use assessment	1	100	100	0	
Tobacco use and help with quitting	7	71	57	43	
Average		83	73	55	

*Data Elements Required for Measure do not include the adolescent's age.

of the chlamydia denominator identified 47.7 percent of patients in the denominator, whereas the computer-identified denominator included 38.2 percent. For the depression denominator, Kappa was 0.17 (95 percent CI = [0.06, 0.27]). The depression screening measure required that adolescents not have an existing depression diagnosis or treatment. The manual denominator was smaller (86.8 percent vs. 94.5 percent for computer-extracted denominator data) because the raters were able to identify more adolescents with existing diagnoses or treatment.

Concurrent Validity Test: Agreement between Computer-Extracted and Manual Data

Table 4 reports the Kappa-measured agreement between the computer-extracted and manual data. The overall agreement was a low-to-moderate Kappa = 0.36. The albeit arbitrary conventions for interpreting Cohen's Kappa treat Kappa ≥ 0.80 as very strong agreement (Landis and Koch 1977), and meningococcal vaccination and Tdap/Td vaccination exceed this mark. Agreement was significantly greater at site 1 than at sites 2 or 3. Table 4 also

Table 4: Agreement between Performance Rates Based on Manual Reviews versus Computer Extracts by Site: Kappa, Sensitivity, and Specificity

Measure	Site 1			Site 2			Site 3			All Sites		
	K	Sensitivity (%)	Specificity (%)	K	Sensitivity (%)	Specificity (%)	K	Sensitivity (%)	Specificity (%)	K	Sensitivity (%)	Specificity (%)
Immunizations domain												
HPV vaccination	0.78	77.8	100.0	0.00*	0.0	100.0	~ [†]	~ [†]	100.0	0.15	17.9	100.0
Meningococcal vaccination	1.00	100.0	100.0	0.93	98.6	100.0	~ [†]	~ [†]	100.0	0.98	98.8	100.0
Tetanus, diphtheria, and pertussis (Tdap/Td) vaccination	1.00	100.0	100.0	0.82	98.7	83.3	~ [†]	~ [†]	100.0	0.96	98.9	97.5
All measures in domain	0.95	93.9	100.0	0.50	82.2	95.8	~ [†]	~ [†]	100.0	0.78	84.0	99.1
Screenings domain												
Blood pressure screening	0.00*	0.0	100.0	0.00*	0.0	100.0	0.00*	0.0	100.0	0.00*	0.0	100.0
Chlamydia screening	0.90	94.9	95.7	0.83	86.8	100.0	0.67	64.3	100.0	0.77	78.2	99.3
Depression screening	0.42	31.7	100.0	0.00*	0.0	100.0	0.62	51.4	100.0	0.38	28.2	100.0
All measures in domain	0.65	54.9	99.7	0.40	30.6	100.0	0.67	56.8	100.0	0.58	47.1	99.9
Counseling and risky behavior assessments domain												
Nutrition counseling	0.11	9.8	100.0	0.00	0.0	100.0	0.46	42.9	100.0	0.08	9.7	100.0
Physical activity counseling	0.00*	0.0	100.0	0.00*	0.0	100.0	0.01	8.9	92.5	0.01	2.3	98.3
Sexual activity documentation	0.67	74.8	98.6	0.30	53.4	100.0	0.04	59.3	100.0	0.37	61.4	99.1

continued

Table 4. Continued

Measure	Site 1			Site 2			Site 3			All Sites		
	K	Sensitivity (%)	Specificity (%)	K	Sensitivity (%)	Specificity (%)	K	Sensitivity (%)	Specificity (%)	K	Sensitivity (%)	Specificity (%)
Alcohol use assessment	0.72	73.6	97.2	0.00*	0.0	100.0	0.00*	0.6	100.0	0.15	19.3	98.8
Substance use assessment	0.61	61.2	99.0	0.00*	0.0	100.0	0.00*	0.0	100.0	0.08	14.0	99.4
Tobacco use and help with quitting	0.15	29.2	84.7	0.74	81.1	94.4	0.41	81.5	72.4	0.52	68.8	86.9
All measures in domain	0.46	49.4	95.9	0.16	23.0	98.7	0.13	34.3	93.8	0.24	33.9	96.5
All measures	0.54	52.5	97.5	0.29	33.9	99.2	0.30	37.0	98.2	0.36	39.5	98.3

Note. “Performance across Measures” means that data were pooled across measures to calculate performance. Agreement at site 1 was higher than at either site 2 or site 3: Kappa (site 1)—Kappa (site 2) = 0.25 (95% CI = [0.19, 0.31]) and Kappa (site 1)—Kappa (site 3) = 0.23 (95% CI = [0.17, 0.30]). Kappas for sites 2 and 3 were much closer: Kappa (site 2)—Kappa (site 3) = −0.01 (95% CI = [−0.06, 0.04]). The differences among the measurement-domain Kappas are all large: Kappa (immunizations)—Kappa (screenings) = 0.20 (95% CI = [0.12, 0.29]), Kappa (immunizations)—Kappa (risk assessments) = 0.54 (95% CI = [0.48, 0.30]), and Kappa (screenings)—Kappa (risk assessments) = 0.34 (95% CI = [0.29, 0.39]). Site 1 had higher sensitivity than site 2; delta = 18.6% (95% CI = [12.9%, 24.6%]) and was also higher than site 3 (delta = 15.5%, 95% CI = [9.9%, 21.6%]); but sensitivity was more similar at sites 2 and 3 (delta = −3.1%, 95% CI = [−7.1%, 0.7%]). Immunizations had higher sensitivity than screenings: delta = 36.9% (95% CI = [30.0%, 43.7%]) and was also higher than risk assessments (delta = 50.2%, 95% CI = [45.5%, 55.0%]); and similarly screenings had higher sensitivity than risk assessments (delta = 13.2%, 95% CI = [7.9%, 18.6%]). Specificity does not vary significantly as a function of site or measurement domain.

*Kappa is at chance level (0.00) because no services were detected in the computer-extracted data.

†Kappa values could not be calculated because no services were detected in either manual or computer-extracted data extraction. Immunization data were missing for site 3 because this was a gynecology clinic that did not perform vaccinations.

‡Sensitivity values could not be calculated because no services were detected either in the manual data extraction or computer-extracted data extraction.

shows that computer-extracted data have high specificity but low to moderate sensitivity. The differences across sites in specificity are only about 1 percent. Sensitivity, however, varied substantially across sites and measurement domains. Sensitivity was greater at site 1 than at sites 2 or 3. Sensitivity for the immunization measures was greater than for the screening measures, and sensitivity for the screening measures was greater than for the risk assessment/counseling measures.

Bias: Relative Performance on Measures Calculated from Manual Reviews and Computer Extracts

Table 5 presents the performance rates calculated from manual and computer-extracted data, for each measure, by site, and for all sites. The Relative Performance of measures based on computer-extracted compared with results based on manual data was 41.1 percent (95 percent CI = [38.9, 43.2]). Performance rates based on computer-extracted data were less than or equal to performance rates based on manual data for every measure at every site (that is, Relative Performance $\leq$ 100 percent; the upper bounds of the 95 percent bootstrap confidence intervals around the Relative Performance was <100 percent for 26 of 36 possible measures by site comparisons). Table 5 shows that Relative Performance was higher at site 1 compared to site 2 or site 3, meaning that site 1's computer-extracted results were closer to the results based on manual review. Finally, the Relative Performance of results based on computer-extracted data in the immunization domain was superior to the Relative Performance in the screenings domain, and the Relative Performance in the screenings domain was superior to the Relative Performance for the risk assessments/counseling domain.

DISCUSSION

We found low-to-moderate agreement between data on care extracted from EHRs by computer and data obtained through manual review. The key problem in agreement was low sensitivity: computer-extracted measures failed to detect documentation of services in the EHRs that were detected by human reviewers. Low specificity—computer-extraction finding service events missed by humans—was not a significant problem. This is consistent with the findings of Hazlehurst and colleagues, who similarly found greater problems with sensitivity than specificity in the quality measures they studied

(Hazelhurst et al. 2012). A likely explanation for low sensitivity is that information about care in the EHR is sometimes recorded in unstructured fields, such as free text or scanned documents, and the standard methods of EHR-based quality measurement cannot access data in unstructured fields. Conversely, other quality data were recorded in structured fields precisely because the EHR has been used either to order or to bill for them, for example, screening tests and immunizations. As a result, performance rates for adolescent well-care estimated using computer-extracted data were often substantially lower than performance rates based on manual data. Similarly, Baker et al. (2007) found that computer-extracted EHR data often missed details critical to quality measurement for heart failure patients and Parsons et al. (2012) found that EHR-derived measures can undercount practice performance compared with the manual review of electronic charts.

This explanation is consistent with the variation of our results across site and measure domain. The proportions of data elements implemented as structured fields were higher at site 1 than sites 2 or 3. This is likely to be one reason why the agreement Kappas and the sensitivities were higher for site 1 than sites 2 or 3, and similarly the Relative Performance of computer-extracted results. To our knowledge there are no previous studies examining whether variation in the results based on computerized versus manually extracted data covaries with the use of structured fields. The finding that agreement and bias in quality measurement vary by site is important given the desire to use quality measures to compare providers and incentivize them based on quality. Future research should continue to examine variation in agreement by site, because such variation raises questions about the use of measures based on computer-extracted data to profile the performance of sites.

Our results also showed better agreement between manually extracted and computerized extracts for immunizations measures compared to screening and risk assessment measures. Again, this corresponds to the relative order for the proportions of data elements represented as structured fields for these measure types.

Should quality measures calculated from computer-extracted data be considered valid? Meningococcal vaccination and Tdap/Td vaccination had very high Kappa agreement and Relative Performance and clearly meet our concurrent validity tests. Performance on the HPV vaccination rate was lower. This was likely due to the stringency of the measure's numerator requirement (i.e., all three doses of the vaccination series had to have been given by the age of 13 years) and that there is less acceptance of this vaccine among the community. Our results suggest, therefore, that quality measures based on

Table 5: Denominator Sizes, Performance Rates, and Relative Performance of Computer-Extracted and Manual Data, by Sites and Quality Measure

Measure	Site 1				Site 2			
	Performance Rates				Performance Rates			
	Denominator	Manual (%)	Computer (%)	Relative Performance	Denominator	Manual (%)	Computer (%)	Relative Performance
Immunizations domain								
HPV vaccination	18	50.0	38.9	77.8	40	75.0	0.0	0.0*
Meningococcal vaccination	37	29.7	29.7	100.0	82	90.2	89.0	98.6
Tetanus, diphtheria, and pertussis (Tdap/Td) vaccination	37	35.1	35.1	100.0	82	92.7	92.7	100.0
All measures in domain	92	35.9	33.7	93.9	204	88.2	73.0	82.8*
Screenings domain								
Blood pressure screening	200	5.5	0.0	0.0*	200	18.0	0.0	0.0*
Chlamydia screening	62	62.9	61.3	97.4	61	62.3	54.1	86.8*
Depression screening	174	23.6	7.5	31.7*	195	17.4	0.0	0.0*
All measures in domain	436	20.9	11.7	56.0*	456	23.7	7.2	30.6*
Counseling and risky behavior assessments domain								
Nutrition counseling	122	41.8	4.1	9.8*	176	81.2	0.0	0.0*
Physical activity counseling	122	45.9	0.0	0.0*	176	58.5	0.0	0.0*
Sexual activity documentation	200	63.5	48.0	75.6*	200	81.5	43.5	53.4*
Alcohol use assessment	200	45.5	35.0	76.9*	200	45.5	0.0	0.0*
Substance use assessment	200	49.0	30.5	62.2*	200	78.5	0.0	0.0*
Tobacco use and help with quitting	200	44.5	21.5	48.3*	200	55.5	47.5	85.6*
All measures in domain	1,044	49.0	26.3	53.7*	1,152	66.7	15.8	23.7*
All measures	1,572	40.5	22.7	56.1*	1,812	58.3	20.1	34.5*

Note. “Denominator” is the count of the patients in the denominator, based on the manually extracted data. For “All Measures in Domain” and “All Measures,” however, “Denominator” is the sum of the counts of patients included in the measures denominators. “Manual (%)” is the performance on the quality measure calculated based on the manual denominator and the manual numerator. “Computer (%)” is the performance on the quality measure calculated based on the manual denominator and the computer-extracted numerator. Relative Performance = $100\% \times \text{Computer Performance Rate} / \text{Manual Performance Rate}$. The * on a Relative Performance value indicates that the upper bound on the 95% bootstrapped confidence interval

Site 3				All Sites			
Performance Rates				Performance Rates			
Denominator	Manual (%)	Computer (%)	Relative Performance	Denominator	Manual (%)	Computer (%)	Relative Performance
9	0.0	0.0	~	67	58.2	10.4	17.9*
10	0.0	0.0	~	129	65.9	65.1	98.8
10	0.0	0.0	~	129	69.0	69.0	100.0
29	0.0	0.0	~	325	65.5	55.4	84.5*
197	3.0	0.0	0.0	597	8.9	0.0	0.0*
162	43.2	27.8	64.3*	285	51.6	40.7	78.9*
149	23.5	12.1	51.4*	518	21.2	6.0	28.2*
508	21.9	12.4	56.8*	1,400	22.1	10.5	47.4*
96	43.8	18.8	42.9*	394	59.9	5.8	9.7*
96	58.3	8.3	14.3*	394	54.6	2.0	3.7*
197	98.5	58.4	59.3*	597	81.1	49.9	61.6*
197	86.8	0.5	0.6*	597	59.1	11.9	20.1*
197	87.8	0.0	0.0*	597	71.7	10.2	14.3*
197	85.3	73.6	86.3*	597	61.6	47.4	76.9*
980	82.0	29.3	35.7*	3,176	65.6	23.4	35.7*
1,517	60.3	23.1	38.3*	4,901	53.2	21.9	41.1*

around Relative Performance was less than 100%. The Relative Performance at site 1 was superior to site 2 (delta = 21.7%, 95% CI = [15.5%, 27.4%]) and was also superior to site 3 (delta = 17.9%, 95% CI = [11.5%, 24.9%]). The Relative Performance at site 2 was similar to site 3 (delta = -3.8%, 95% CI = [-7.5%, -0.2%]). The Relative Performance for the immunization measures was superior to that for the screening measures (delta = 37.1%, 95% CI = [30.5%, 43.8%]) and the risks assessments/counseling measures (delta = 48.8%, 95% CI = [43.8%, 53.6%]). Finally, the Relative Performance of the screenings was superior to the risk assessments (delta = 11.7%, 95% CI = [6.3%, 16.6%]).

computerized extracts may be used to reliably rank the quality of health care organizations on meningococcal vaccination and Tdap/Td vaccination. This finding is not surprising: vaccination has been a central priority for preventive care in children for a long time. As a result, vaccination data are more likely to be captured in a structured field in an EHR and are therefore easily retrievable.

Otherwise, the levels of performance estimated from computer-extracted data often have substantial negative bias: overall Relative Performance <50 percent. It is possible that a systematically biased measure could nevertheless provide reliable information about the rank order of performance at different health care organizations, in the way that a miscalibrated scale will nevertheless accurately report the rank ordering of the weights of different people. However, with only three sites in this study we cannot quantify the reliabilities of rank orderings of sites. Our results do not support the concurrent validity of performance rates calculated from computer-extracted data for the other quality measures examined in this study.

It is clearly possible to do better. Our analyses showed that the low sensitivity of the computer-extracted results was the critical defect in the measures studied here. But not all computer-extracted results have low sensitivity. In a study of the quality of cardiovascular care, Persell et al. (2006) found a relatively high level of sensitivity (82–98 percent) in a study of the correspondence between automatically measured EHR data and data on quality from review of free text. Our results suggest that the validity of quality measures could be improved if more data elements were captured as structured fields. It is unfortunate that in spite of the decades that EHRs have been in existence, clinicians using EHRs must still record information about care in unstructured formats. EHR vendors should design systems so that structured fields can be more readily implemented into EHRs. Similarly, health care systems should invest more tailoring in EHR applications to allow clinicians to incorporate structured data efficiently in their workflows. However, although these efforts will make it easier to capture data in structured fields, improving the documentation of data elements for quality measures will also require that clinicians and other staff prioritize such documentation and change their clinical practices to accomplish it.

However, it is unlikely that this is the only problem. The variation among sites in the proportions of data elements implemented as structured fields were not as striking as the variation in agreement, so there must be more to the story. The rest of the story, we speculate, is that having structured fields in an EHR is necessary but insufficient to create a data set that will support a valid performance rate. To calculate valid quality measure results, clinicians must use these fields to document the care they provide. This study addressed

only one factor that might help explain variation in the validity of quality measures: the use of structured fields in EHRs. It is very likely, however, that other aspects of health information technology and clinical practice affect the validity of quality measures. Multiple workflow and technical factors affect the availability of structured, standardized data for quality measurement. Is the information needed for quality measurement generated during routine care within a clinical practice or as defined by an evidence-based clinical best practice? Is this information captured electronically in the EHR system or another type of health IT system? Can the information be extracted from the EHR database without manual review and, if so, in a format that can be used for automatically calculating and reporting the results of quality measures? Although an EHR system may have the technical function needed to capture, calculate, and report a quality measure's data elements, local workflow and documentation habits may ultimately dictate whether quality data are available and usable from the EHR system. The ability to automatically extract data from EHR databases varies across EHR products and the site personnel's knowledge of the data in their EHRs. Future research on the validity of quality measures should include detailed information about clinical workflow, documentation, and processes of automated EHR extraction.

In summary, it appears that computers often cannot extract quality data from an EHR because providers have placed these data in free text notes. Sometimes this may be because the EHR has not provided a structured field in which to place the data. It is not clear, however, that putting everything in structured fields will solve the problem, because many providers believe that they click too many buttons. Free text, moreover, plays an essential role in communicating information to other providers. A complete solution would make data from both structured fields and free text notes available to quality measurement. Such an approach would involve the use of text searching or natural language programming techniques to enable machines to understand clinicians' free text entries and accurately extract quality data from them.

Limitations

The representativeness of our findings is limited by having studied only three sites and three measurement domains. Similarly, although we have studied more quality measures than many previous studies, they are all from a single area: adolescent well-care.

The use of manually extracted data from an EHR system as a gold standard is also a limitation, because services may have been delivered to the

adolescent that were not entered into the EHR in any form. Care may therefore have been better than is indicated by either the manual or computer-extracted data, so agreement between manual and computer-extracted data may not be sufficient evidence for the validity of the computer-extracted measure. However, lack of agreement between manual and computer-extracted data raises questions about the validity of the computer-extracted results in the context of current clinical practice.

Finally, it is possible that some of the low sensitivity of measures calculated from computer-extracted data could be attributed to errors made by sites in programming the computer extractions. We provided sites with specifications of the measures and asked personnel at the sites to write the relevant code. These errors could have occurred either in the configuration of the EHR product or the programming of the data extraction. Similarly, we were unable to directly inspect the EHRs at each site to determine which data elements were in structured fields. Instead, we relied on survey reports from site personnel, which in some cases may have been in error.

Conclusion

These data support the concurrent validity of computer-extracted quality measures for immunizations. They raise questions, however, about using computer-extracted data to assess the provision of screening or risk assessments/counseling for adolescents. Agreement between manually and computer-extracted quality measures is variable, with many computer-extracted measures indicating substantially lower quality of care. The accuracy of quality reporting through computerized EHR extraction depends on the presence and use of structured data fields, with the highest agreement found in the site and domain of measures that had the greatest concentration of structured fields. EHR vendors and health care providers should work together to improve documentation of care in ways that are well adapted to practice workflow, such that computerized data extraction from EHRs can result in valid measures of clinical performance.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This project was supported by grant number U18HS020503 from the Agency for Healthcare Research and Quality (AHRQ). The content is solely the responsibility of the authors and

does not necessarily represent the official views of the Agency for Healthcare Research and Quality. Suzanne Morton, Sepheen C. Byron, Aldo Tinoco, Vivian Kong, and Sarah Hudson Scholle are employed by the National Committee for Quality Assurance (NCQA), a not-for-profit organization that develops and maintains quality measures. The remaining authors report no conflicts of interest with respect to this work.

Disclosures: None.

Disclaimers: None.

REFERENCES

- Baker, D. W., S. D. Persell, J. A. Thompson, N. S. Soman, K. M. Burgner, D. Liss, and K. S. Kmetik. 2007. "Automated Review of Electronic Health Records to Assess Quality of Care for Outpatients with Heart Failure." *Annals of Internal Medicine* 146 (4): 270–7.
- Fuhlbrigge, A., V. J. Carey, J. A. Finkelstein, P. Lozano, T. S. Inui, and K. B. Weiss. 2008. "Are Performance Measures Based on Automated Medical Records Valid for Physician/Practice Profiling of Asthma Care?" *Medical Care* 46 (6): 620–6.
- Gold, R., H. Angier, R. Mangione-Smith, C. Gallia, P. J. McIntire, S. Cowburn, C. Tillotson, and J. E. DeVoe. 2012. "Feasibility of Evaluating the CHIPRA Care Quality Measures in Electronic Health Record Data." *Pediatrics* 130 (1): 139–49. doi:10.1542/peds.2011-3705.
- Hazelhurst, B., M. A. McBurnie, R. A. Mularski, J. E. Puro, and S. L. Chauvie. 2012. "Automating Care Quality Measurement with Health Information Technology." *The American Journal of Managed Care* 166 (6): 313–9.
- Iezzoni, L. I. 1997. "Assessing Quality Using Administrative Data." *Annals of Internal Medicine* 127 (8 Pt 2): 666–74.
- Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33: 159–74.
- Linder, J. A., E. O. Kaleba, and K. S. Kmetik. 2009. "Using Electronic Health Records to Measure Physician Performance for Acute Conditions in Primary Care." *Medical Care* 47 (2): 208–16.
- MacLean, C. H., R. Louie, P. G. Shekelle, C. P. Roth, D. Saliba, T. Higashi, J. Adams, J. T. Chang, C. J. Kamberg, D. H. Solomon, R. T. Young, and N. S. Wenger. 2006. "Comparison of Administrative Data and Medical Records to Measure the Quality of Medical Care Provided to Vulnerable Older Patients." *Medical Care* 44 (2): 141–8.
- Parsons, A., C. McCullough, J. Wang, and S. Shih. 2012. "Validity of Electronic Health Record-derived Quality Measurement for Performance Monitoring." *Journal of the American Medical Informatics Association* 19 (4): 604–9. doi:10.1136/amia-jnl-2011-000557.

- Persell, S. D., J. M. Wright, J. A. Thompson, K. S. Kmetik, and D. W. Baker. 2006. "Assessing the Validity of National Quality Measures for Coronary Artery Disease Using an Electronic Health Record." *Archives of Internal Medicine* 166 (20): 2272–7.
- Roth, C. P., Y.-W. Lim, J. M. Pevnick, S. M. Asch, and E. A. McGlynn. 2009. "The Challenge of Measuring Quality of Care from the Electronic Health Record." *American Journal of Medical Quality* 24 (5): 385–94.
- Tang, P. C., M. Ralston, M. F. Arrigotti, L. Qureshi, and J. Graham. 2007. "Comparison of Methodologies for Calculating Quality Measures Based on Administrative Data versus Clinical Data from an Electronic Health Record System: Implications for Performance Measures." *Journal of the American Medical Informatics Association* 14 (1): 10–5.
- U.S. Department of Health and Human Services. 2011. "Report to Congress: National Strategy for Quality Improvement in Health Care." Washington, DC [accessed on January 6, 2014]. Available at: <http://www.chhs.ca.gov/Documents/LGHCTF%20National%20Quality%20Strategy.pdf>
- U.S. Department of Health and Human Services. 2012. "Meaningful Use Stage 2 – Clinical Quality Measures." Meaningful Use – Stage 2 [accessed on February 26, 2013]. Available at <http://www.healthit.gov/policy-researchers-implementers/clinical-quality-measures>
- Wasserman, R. C. 2011. "Electronic Medical Records (EMRs), Epidemiology, and Epistemology: Reflections on EMRs and Future Pediatric Clinical Research." *Academic Pediatrics* 11 (4): 280–7. doi:10.1016/j.acap.2011.02.007.
- Weiner, W., T. E. Stump, C. M. Callahan, J. N. Lewis, and C. J. McDonald. 2005. "Pursuing Integration of Performance Measures into Electronic Medical Records: Beta-adrenergic Receptor Antagonist Medications." *Quality and Safety in Health Care* 14 (2): 99–106.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Appendix SA2: Using Automated Extracts from Electronic Health Records to Measure the Quality of Adolescent Well-Care.