The challenges of linking health insurer claims with electronic medical records

Suzanne L. West
RTI International, USA; University of North Carolina at Chapel Hill, USA

William Johnson
Arizona State University, USA

Wendy Visscher and Marianne Kluckman
RTI International, USA

Yue Qin
Arizona State University, USA

Ann Larsen
RTI International, USA

Abstract
This article explores the challenges inherent in linking data from disparate sources—electronic medical records (EMR) and health insurer claims—and the probable benefits of doing so to evaluate several quality measures associated with diabetes. Using the business associate agreement provision of the Health Insurance Portability and Accountability Act, we were able to link health insurer claims with EMR data; however, when restricting the linked data to patients with at least one medication and one diagnosis in the evaluation year, we lost 90 percent of our linked population. Whether this loss was due to difficulties in extracting the data from site EMRs, to changes in insurer coverage over time, or to both was not discernible. Because linking EMR data to health insurer claims can produce a clinically rich longitudinal data set, assessing the completeness and quality of the data is critical to health services research and health-care quality measurements.

Keywords
Electronic medical records, EMRs, Health Information Technology for Economic and Clinical Health Act (HITECH), Health Insurance Portability and Accountability Act (HIPAA), health insurer claims, linkage, Patient Protection and Affordable Care Act (PPACA)

Corresponding author:
Suzanne L. West, RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709-2194, USA. Email: swest@rti.org
Introduction

Two important sources of electronic data that may be used for health services research and quality-of-care evaluations are health insurer claims (claims) and electronic medical records (EMRs). While claims have been used for decades for both kinds of research, EMRs have been available for less time, but due to their level of detail, they are gaining in use. Each data source has its unique value, but linking the two sources, although fraught with legal challenges and technological difficulties, may prove synergistically useful for research and policy applications. To date, we know of no study that has linked EMR data to health insurer claims from two disparate data sources. Herein, we describe a case study that uses a de-identified, granular, longitudinal data set consisting of linked claims and EMR data for evaluating diabetes quality of care.

EMRs

According to the Office of the National Coordinator for Health Information Technology, an EMR is “a real-time patient health record with access to evidence-based decision support tools that can be used to aid clinicians in decision-making.” EMR data elements vary from system to system; however, the following are almost always captured as discrete data and are required for certified ambulatory systems: age, gender, diagnosis, medical history, medications prescribed, lab and procedure orders and results, allergies, immunizations, and vital signs.

Enacted as part of the American Recovery and Reinvestment Act of 2009, the Health Information Technology for Economic and Clinical Health Act (HITECH Act) fostered physician and hospital adoption of certified EMRs, initially in 2011 by providing reimbursement incentives. Starting in 2015, however, providers caring for Medicare beneficiaries who do not adopt certified EMRs will be penalized. Consequently, an increasing amount of EMR data will be available for patient-centered outcomes research, drug and device safety evaluations, and assessment of the quality and continuity of patient care.

Comprehensiveness of a patient’s medical and treatment history is essential to effective medical care, but it is often unavailable because of the current fragmentation of the US health-care system. Patients see many clinicians, many of whom are not part of integrated health systems that share EMRs across all specialties and affiliated hospitals. If the patient is being cared for by independent clinicians (i.e. clinicians not in an integrated system), each of whom has a separate EMR, the different EMRs may not be interoperable (i.e. unable to seamlessly share patient data). If EMR data are combined with other health and registry data as part of a health information exchange (HIE), linkage can facilitate sharing of clinical information across providers to minimize gaps in care. Currently, many between-organization HIEs are under development, but few are fully operational (i.e. actually transmitting data useful for health-care stakeholders).

Health insurer claims data

Claims data provide information about every service, from any medical provider, paid for by the patient’s health insurer. For more than 20 years, because of the relative completeness of these data, researchers have used claims data to study drug safety and the economics of health care. However, claims data have limitations, especially that they lack granular clinical information such as test results. Also, private health-care insurance coverage is generally employer based, so when employers change plans or employees change companies, the longitudinal nature of health-care information is lost. Public insurance data, such as Medicaid claims, are not universally available,
and patients often migrate in and out of the system as their eligibility changes. When individuals are covered by multiple health insurers, all of which use different patient identifiers, linking the claims to create a longitudinal patient record is difficult—and has been made more so by the Health Insurance Portability and Accountability Act (HIPAA) of 1996 and its expansion under the HITECH Act, which mandate that health-care data used for research be purged of specific identifiers, such as names, medical record numbers, and Social Security numbers.

**HIPAA and its influence on research**

The primary goals of HIPAA legislation were to improve patients’ ability to continue their health insurance, regardless of their employment status; reduce fraud; and simplify the administration of health coverage. The HIPAA Privacy Rule, developed under administrative simplification provisions, required the adoption of standards to protect individually identifiable health information. This protected health information (PHI) refers to information held by a “covered entity” (health provider, health plan, or health-care clearinghouse) related to a person’s health or treatment and containing any of 18 specific identifiers (for a list, see Appendix 1). The Privacy Rule states that a covered entity may not use or disclose PHI except as permitted or required by this legislation. As part of the original HIPAA statute, section 1173, the Secretary of the Department of Health and Human Services was tasked with adopting standards for a unique health identifier for each individual. The development of a unique patient identifier has not occurred and may never occur. (For a more detailed description of HIPAA and use of health-care data for research, see Appendix 1.)

In 2007, the Institute of Medicine (IOM) convened a committee to examine the effect of the HIPAA Privacy Rule on research and to recommend other approaches for facilitating health research while maintaining the privacy of health information. The committee found wide variability in the way that the research provisions of the HIPAA Privacy Rule were interpreted and implemented by individual covered entities. Because of this variability, some researchers reported to the committee that some of their studies were delayed or simply could not be conducted. At a minimum, to reduce this variability, the committee recommended that the Department of Health and Human Services either revise the research provisions, or provide more guidance to covered entities. However, the committee’s foremost recommendation was to adopt an entirely new framework that would exempt health research from HIPAA and instead impose other oversight mechanisms for safeguarding PHI. Now, 5 years later, the IOM’s recommendations have not been adopted, and the restrictions on the use of PHI for research have become even more stringent under the HITECH Act.

**Linkage of health insurer claims and EMR data**

The difficulty of linking a patient’s complete medical information (EMR and claims data) across providers and health plans over a multiyear period has limited researchers’ ability to conduct longitudinal studies. If possible, such linkage would broaden, appreciably, the research questions that could be evaluated with use of either EMR or claims data individually. For example, using non-linked ambulatory EMR data to conduct a diabetes study, one could evaluate the comparative effectiveness of diabetes medications on glycosylated hemoglobin over time, determining which medication maintains glucose control longest and with the minimum number of side effects. However, these ambulatory EMR data could not be used to determine what proportion of diabetics using specific medications had an inpatient stay with outcomes such as myocardial infarction. Longitudinal data about hospitalizations would not be available except from patients themselves.
Linkage of the EMR with claims data from patients’ health insurers would overcome such difficulties and would permit analysis of both research questions.

**Methods**

We set out to identify a source of ambulatory EMR data that had undergone some processing to standardize coding schemes across providers and that had access to patient identifiers that could be used for direct linkage to claims data. The EMR data source we identified was derived from a commercial vendor’s clinical data warehouse with data from 2007 forward, which contained data for more than 6 million lives from approximately 37 medical groups; all the medical groups gave permission for their data to be used for research.

Similarly, we sought a source of claims data that would include multiple insurers, would provide sufficient overlap in calendar time and patient population with the EMR data, and would also have patient identifiers to allow linkage with the EMR data. Developed by researchers at Arizona State University (ASU), Arizona HealthQuery (AZHQ) is a data system that contains data from 1998 onward, from a variety of sources and payers, for several million individuals, to help state administrators, hospitals, health insurers, and other health-care providers assess the health status and health-care needs of the state. Because AZHQ contained data for millions of Arizonans, and the ASU researchers were fully aware of the HIPAA requirements for use of identifiable data, we decided to use AZHQ claims data for the study.

Typically, when conducting research with covered entities’ data that include any of the 18 HIPAA identifiers (i.e. PHI), the researcher must obtain individual consent (or authorization) from the patients. If doing so is unfeasible, the HIPAA Privacy Rule provides two other research options. One is for the researcher to request a waiver of authorization from an institutional review board (IRB) or a privacy board. Another option is for the researcher to request from the covered entities a limited data set that excludes the 16 direct identifiers included in the list of 18 HIPAA identifiers (for details, see Appendix 1). For us, obtaining individual authorization from all the patients for whom we needed data was unfeasible. Likewise, use of limited data sets was impractical because we needed direct identifiers to link the EMR with claims data.

To maintain compliance with the HIPAA Privacy Rule, we had to engage a trusted third party to link the AZHQ personal identifiers with those from the commercial vendor’s EMR data. Because the AZHQ staff would not be conducting any analyses on the merged EMR-claims data, they could and did function as the trusted third party. To do so, however, AZHQ had to enter into a HIPAA business associate agreement (BAA) with the commercial EMR vendor. In addition, the ASU IRB granted a HIPAA waiver attesting that obtainment of individual patient consent would be impracticable and required ASU investigators to affirm that they would be careful custodians of the data to protect patient privacy. Because RTI International investigators would not have actual patient identifiers (AZHQ replaced each with an anonymized identifier) and would be unable to link the data back to individual patients, RTI’s IRB granted an exemption for the study under 45 C.F.R. § 46.101(b).

In addition to negotiating a BAA and obtaining IRB approvals, ASU had to adhere to its own procedures regarding data providers. Investigators who want to use AZHQ data for research must obtain approvals from the health insurers whose patients’ data would be part of the study. ASU received approval from four health plans; approval was denied by a fifth plan.

Once all the preliminary approvals were obtained and the analysis plan was established, the commercial EMR vendor began delivery of data from two Arizona medical groups. AZHQ worked with the EMR vendor to quality control the data sent to it via a secure file transfer protocol (FTP) transmission. To link the EMR and claims data sets, AZHQ used an exact-match algorithm using
patient names, Social Security numbers, and birth dates. RTI received an encrypted, linked claims-EMR file containing fully de-identified data (i.e. all 18 HIPAA identifiers were removed) with which to conduct analyses of diabetes care quality.

**Results**

Identifying the potential data partners, obtaining the necessary approvals for use of the EMR data provider and the claims data from the plans, negotiating a HIPAA BAA between the EMR data provider and the trusted third party (AZHQ), and obtaining IRB approval at both ASU and RTI took nearly 12 months. Unfortunately, no one noticed until just prior to uploading the data to RTI that the BAA and ASU’s IRB approval required that the linked data set be completely anonymized, which meant deleting all dates of service before the analytic files were sent to RTI. Thus, RTI received a linked data set that contained no dates but, instead, patient ages derived from the dates. Because we were trying to match events by exact occurrence date, ASU converted each date to an age carried out to the second decimal place.

The EMR vendor extracted, transformed, and downloaded data from two large multispecialty medical sites in Arizona into its data warehouse. It then identified patients with diabetes (ICD9 code 250.xx), extracted data files from its data warehouse for this cohort, and then sent these password-protected, encrypted EMR data files via secure FTP server to AZHQ. This EMR cohort served as the baseline for linkage with the AZHQ claims data.

The EMR vendor sent AZHQ 12 files (260 megabytes) for 18,048 patients. Of these, 12,075 were linked to AZHQ claims data (67%). The nearly 6000-person discrepancy between the EMR and claims data occurred, in large part, because a fifth Arizona health insurer did not approve use of their data for the linkage study. The process of linkage based on patient identifiers is diagrammed in Figure 1.

Because this project was the EMR vendor’s first attempt at extracting and preparing a data set for outside use, the learning curve was steep for the entire project team. EMR data were downloaded iteratively over a 4-month period, with each download requiring an evaluation to ensure that the data were being extracted from the medical practices in a form usable for research. AZHQ’s initial quality control procedures involved evaluating whether patients had data for each of the data files. For example, in the first upload of EMR data to AZHQ, all of the patients had at least one medication and one diagnosis, 95 percent had vital signs such as age and weight, 12 percent had a blood pressure measurement, 68.3 percent had a lab test, but only 7.8 percent had a code indicating that a procedure had been done in the physician’s office. More importantly, however, the dates for the events (diagnoses, medications, and lab tests) appeared to be time-stamped according to when the EMR information was last updated, not when the event actually occurred. After the problems with the download were identified, the EMR vendor implemented a new data model for data extraction, and many of these problems resolved on the next data upload to AZHQ.

The software changes that the vendor had made to enhance clinical workflow affected what data were available for extraction and where patient data might be located within the EMR data tables. Although these changes posed no problem for clinical management of patients, they proved a challenge during extraction of the data from the medical sites. In particular, the encounter identifier is an important variable for researchers because it connects all the events that occur during a patient visit: the vitals that are taken, the diagnoses made, the medications prescribed, and the lab tests that are done. The two clinical sites providing data for the study had more than one version of the EMR software, but only the latest version incorporated an encounter number. To overcome this obstacle, we used an algorithm comprising a variable that indicated that the patient “arrived” for the visit and the date for that visit. This algorithm allowed us to identify all events that occurred on that date.
ASU requested use of claims data from 5 health plans, 4 gave approval

1,789 patients analyzed

A new study identifier is generated for each patient in the linked dataset

All date variables converted to patient’s age (to two decimal places) at which event occurred

De-identified claims and EMR data encrypted and sent to RTI via secure file transfer protocol

- Diagnosed with DM earlier than 2007
- At least one diagnosis
- At least one medication in both EMR and claims data in 2008

Figure 1. Electronic health records and claims linkage process, focusing on diabetes mellitus patients. ASU: Arizona State University; EMR: electronic medical record.
Similarly, to identify duplicate EMR records, another algorithm was derived consisting of the patient identifier, the record type, table type, and version identifier.

According to our initial descriptive analyses, some of the diabetic patients in the claims data set did not have medication claims available for analysis (Table 1). Likewise, some of the patients in the EMR data set did not have prescription data. We were also missing data on diagnoses from both the claims and EMR data sets. To overcome these missing data challenges, we restricted our analysis to patients who were diagnosed with diabetes before or during 2007 and who had at least one medication and at least one diagnosis in both the EMR and the claims data in 2008. This restriction reduced our final study population from 12,075 to 1178 patients (Table 2).

Table 2 shows the demographics of the EMR, claims, and linked populations that RTI received from AZHQ. We also present a comparison with the Arizona population from the 2010 Census. Although we were unable to obtain the race/ethnicity distribution for the EMR population for comparison with the 2010 Census, the race/ethnicity distribution of the claims population is similar to the Census figures. Once we restricted the population by diagnosis and medication, the age, gender, and race/ethnicity distribution shifted significantly.

Table 3 provides the percent agreement and \( \kappa \), comparing the EMR and claims data for “ever use” of at least one diabetes medication as indicated by active, complete, or discontinued medication status in the EMR in 2008. Overall, agreement was good, ranging from 80.3 to 92.6 percent, with \( \kappa \)s in the moderate range. However, the nonconcordance indicates that neither the EMR nor the claims suffice for illuminating the complexity of medication exposure.

Table 4 evaluates the timing between the EMR prescription and the claims dispensing for the first diabetes medication prescribed in 2008, using the population of patients who had medications in both the EMR and claims files. For metformin, insulin, and sulfonylureas, most of the medications were dispensed within 1 week of their prescription. Patients were less likely to fill their thiazolidinediones prescriptions within 1 week (34.2%) and were less likely to fill them at any time (48.1%). The fact that none of the thiazolidinediones are available as generics and are very costly without medication insurance coverage may explain why many were not dispensed within 365 days of being prescribed. We also observed that 20–40 percent of diabetes medications prescribed according to an EMR were not dispensed within the subsequent 365 days, according to claims.

We evaluated the agreement between the EMR and claims data for identification of comorbidities associated with diabetes and the laboratory tests typically done in these patients. Both hypertension (agreement = 70%, \( \kappa = 0.39 \)) and obesity (agreement = 85.3%, \( \kappa = 0.42 \)) were documented in both data sources. Table 5 shows that HbA\textsubscript{1c} and lipids were better documented in claims than in the EMR, but the discrepancy for lipids was less than that for HbA\textsubscript{1c}.

### Table 1. The comparative availability of medication, diagnosis, and lab data from the EMR and claims databases, 2008 data.

<table>
<thead>
<tr>
<th>Type of data missing</th>
<th>EMR</th>
<th>Claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medications</td>
<td>6072 (33.6%)</td>
<td>9682 (79.4%)</td>
</tr>
<tr>
<td>Diagnoses</td>
<td>8986 (50.2%)</td>
<td>5285 (43.4%)</td>
</tr>
<tr>
<td>Lab</td>
<td>17,122 (94.9%)</td>
<td>4812 (39.5%)</td>
</tr>
<tr>
<td>Total</td>
<td>18,048</td>
<td>12,189</td>
</tr>
</tbody>
</table>

EMR: electronic medical record.
Missing indicates that no records for this type of data were in the file.
This unique project linked data from two disparate electronic data sources: claims data that provide a broad view of most, if not all, of the care provided to a patient, and EMR data, which are more granular. EMR data may be less comprehensive than claims if derived from only one of the clinical practices from which a patient seeks care, unless the practice is part of an integrated delivery system that is providing the patient with all of his or her ambulatory and inpatient care. However, only

Table 2. Demographic information for patients in the EMR, claims, and linked files as compared with the total Census population for Arizona.

<table>
<thead>
<tr>
<th></th>
<th>EMR data (N = 18,048)</th>
<th>Claims data (n = 12,189)</th>
<th>Linked data (n = 1178)</th>
<th>Arizona 2010 Census data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diabetes patients</td>
<td>Diabetes patients</td>
<td>Diabetes patients</td>
<td></td>
</tr>
<tr>
<td>Age in 2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>62.1</td>
<td>61.9</td>
<td>56.8</td>
<td>35.9b</td>
</tr>
<tr>
<td>Range</td>
<td>17–109</td>
<td>17–104</td>
<td>17–98</td>
<td></td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>9283 (51.4)</td>
<td>6593 (54.1)</td>
<td>792 (67.2)</td>
<td>3,216,194 (50.3)</td>
</tr>
<tr>
<td>Male</td>
<td>8764 (48.6)c</td>
<td>5596 (45.9)</td>
<td>386 (32.8)</td>
<td>3,175,823 (49.7)</td>
</tr>
<tr>
<td>Race, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>NA</td>
<td>7802 (67.2)</td>
<td>532 (45.4)</td>
<td>57.8</td>
</tr>
<tr>
<td>Black</td>
<td>NA</td>
<td>417 (3.6)</td>
<td>80 (6.8)</td>
<td>3.7</td>
</tr>
<tr>
<td>Hispanic</td>
<td>NA</td>
<td>2785 (24.0)</td>
<td>504 (43.0)</td>
<td>29.6</td>
</tr>
<tr>
<td>Native American</td>
<td>NA</td>
<td>255 (2.2)</td>
<td>18 (1.6)</td>
<td>4.0</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>NA</td>
<td>154 (1.3)</td>
<td>28 (2.4)</td>
<td>2.8</td>
</tr>
<tr>
<td>Other</td>
<td>NA</td>
<td>204 (1.8)</td>
<td>10 (0.9)</td>
<td>2.1</td>
</tr>
</tbody>
</table>

EMR: electronic medical record.

Restricting to those patients who were diagnosed with diabetes before 2007 and had at least one medication and diagnosis in both EMR and claims in 2008.

Median age was based on the full age range, whereas the linkage study focused only on adults.

Because of missing data, numbers do not sum to total. Census categorizes Hispanic ethnicity as separate.

Table 3. Agreement between the individual’s EMR and claims records for diabetes medications, using only 2008 as the time window of observation.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Use according to EMR and claims</th>
<th>No use according to EMR and claims</th>
<th>Use according to EMR, no use according to claims</th>
<th>No use according to EMR, use according to claims</th>
<th>Percent agreement (κ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metformin</td>
<td>388</td>
<td>558</td>
<td>109</td>
<td>123</td>
<td>80.3 (0.60)</td>
</tr>
<tr>
<td>Insulin</td>
<td>216</td>
<td>766</td>
<td>83</td>
<td>111</td>
<td>83.5 (0.58)</td>
</tr>
<tr>
<td>Sulfonylurea</td>
<td>229</td>
<td>762</td>
<td>106</td>
<td>81</td>
<td>84.1 (0.60)</td>
</tr>
<tr>
<td>Thiazolidinediones</td>
<td>60</td>
<td>1031</td>
<td>48</td>
<td>39</td>
<td>92.6 (0.54)</td>
</tr>
</tbody>
</table>

EMR: electronic medical record.

Data were restricted to those patients who were diagnosed with diabetes before 2007 and had at least one medication and diagnosis in both EMR and claims in 2008.

Discussion

This unique project linked data from two disparate electronic data sources: claims data that provide a broad view of most, if not all, of the care provided to a patient, and EMR data, which are more granular. EMR data may be less comprehensive than claims if derived from only one of the clinical practices from which a patient seeks care, unless the practice is part of an integrated delivery system that is providing the patient with all of his or her ambulatory and inpatient care. However, only

D...
if a patient is receiving care from an integrated delivery system that is a health insurer as well, such as Kaiser Permanente, will there be a guarantee of comprehensive capture of all medical care. Partly because of the complexity of the EMR software and because of the difficulty in extracting information, we were unable to extract, transform, and load all relevant EMR data for this analysis. We believe these difficulties explain our starting with approximately 12,000 linked records but, when restricting them to the most complete data (both drugs and diagnoses in both files), ending with only 1178 patients’ records.

The use or disclosure of PHI for health research is permitted without authorization from the patients, under the Privacy Rule’s research provisions. Most researchers are aware of the limited-data-set research provision under the HIPAA Privacy Rule, which allows access to covered entities’ data

<table>
<thead>
<tr>
<th>Table 4. Comparison of time between the first diabetes medication prescription in the EMR in 2008 and a dispensing for that medication according to claims.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days between prescribing and dispensing</td>
</tr>
<tr>
<td>Metformin (n = 497)</td>
</tr>
<tr>
<td>n (%)</td>
</tr>
<tr>
<td>0–1</td>
</tr>
<tr>
<td>2–7</td>
</tr>
<tr>
<td>8–30</td>
</tr>
<tr>
<td>31–90</td>
</tr>
<tr>
<td>91–365</td>
</tr>
<tr>
<td>No dispensing claim within 1 year</td>
</tr>
</tbody>
</table>

EMR: electronic medical record.
Data were restricted to those patients who were diagnosed with diabetes before 2007 and had at least one medication and diagnosis in both EMR and claims in 2008 (N = 1178). Patients can appear in more than one column if prescriptions for more than one diabetes medication were written.

<table>
<thead>
<tr>
<th>Table 5. Documentation of laboratory measures in EMRs and claims data, stratified by whether the medical center had electronic laboratory interfaces.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documentation of laboratory measure</td>
</tr>
<tr>
<td>Medical group lacked electronic lab interfaces</td>
</tr>
<tr>
<td>(n = 809)</td>
</tr>
<tr>
<td>Glycosylated hemoglobin (HbA1c)</td>
</tr>
<tr>
<td>According to claims 254 (31.4)</td>
</tr>
<tr>
<td>According to EMR 74 (9.2)</td>
</tr>
<tr>
<td>Total, LDL, or HDL cholesterol</td>
</tr>
<tr>
<td>According to claims 235 (29.1)</td>
</tr>
<tr>
<td>According to EMR 67 (8.3)</td>
</tr>
<tr>
<td>Influenza vaccine</td>
</tr>
<tr>
<td>According to claims 185 (22.9)</td>
</tr>
<tr>
<td>According to EMR 226 (27.9)</td>
</tr>
</tbody>
</table>

EMR: electronic medical record; LDL: low-density lipoproteins; HDL: high-density lipoproteins.
Data were restricted to those patients who were diagnosed with diabetes before 2007 and had at least one medication and diagnosis in both EMR and claims in 2008 (n = 1178).
sets containing PHI (e.g. dates, some geographic information) so long as direct identifiers have been stripped. However, the HIPAA rules governing research use of data sets with “full” PHI (i.e. containing direct identifiers) are much less straightforward. The only HIPAA research provision used for linkage research requires attainment of a waiver of authorization from an IRB or privacy board. Obtaining a waiver of authorization requires a ruling that, even though PHI (including personally identifying information such as name, date of birth, state, and zip code) will be disclosed by the covered entities to the researchers, the research poses “no more than minimal risk to the privacy of the individual,” and, therefore, signed authorization from the individual patients is not required.

RTI worked closely with privacy officers from all three institutions (ASU, RTI, and the EMR vendor) and the IRBs at ASU and RTI over a period of several months to negotiate the terms of the HIPAA waiver, BAA, and IRB approvals. In the end, despite careful review of all documents, RTI had to work with a file in which dates had been converted to the patient’s age when the event occurred. Although we anticipated difficulties from using ages instead of dates, we were able to conduct all the required analyses without major challenges.

Use of BAAs and IRB approval for linking data sets is either relatively new or has not been well documented in the literature. We have identified only one published paper describing a similar HIPAA-compliant approach to addressing patient privacy concerns when using data from four different data providers to evaluate the adverse risks associated with medications used to treat autoimmune disorders.13

We did not anticipate the extent of the missing data we identified in the linked data set: only 10 percent of the data remained after applying diagnosis and medication restrictions. One of the reasons so much data were lost to analysis when requiring both diagnosis and medication data is that the patients did not get medical and/or pharmacy benefits. However, in the remaining data, we did find good agreement for ever use of particular diabetes medications, based on both the percent agreement and $\kappa$.

A significant strength of this project was the ability to compare prescribing as recorded in EMRs and dispensing of medications as recorded in claims data, especially for medications used to treat a chronic condition that requires careful monitoring to maintain glycemic control. The EMR data provides insight into prescribing behavior (i.e. how the clinician is treating the patient), whereas the dispensing information indicates whether the patient is filling the prescription, which is one step closer to medication consumption. That most patients in our study had their diabetes medications dispensed within 1 week of the prescription date is reassuring. Our finding that 25–35 percent of patients did not fill their metformin, insulin, or sulfonylurea medication resembles findings by other researchers.14

The paucity of laboratory data from the EMR, even when clinics had an electronic interface with large laboratories, was disappointing because we expected laboratory data to be the greatest value of EMRs. Possibly, the EMR vendor was unable to extract the totality of lab data available. Whether this problem will endure for the EMR vendor or be resolved with subsequent downloads remains to be seen.

The need for linkage across data sets—claims, EMRs, and registries—is growing exponentially as the availability of data grows. Researchers working on active drug safety surveillance will be concerned that their surveillance population likely contains duplication, which is why linkage is so important. As the Public Health Informatics Institute has stated, “Matching person specific data from multiple sources to produce linked data sets or merged records is fundamental to the quality of data and the integrity of the information being provided by the integrated systems.”

Perhaps an EMR-claims linkage has rarely been attempted in the United States15,16 primarily because health research using PHI must comply with the HIPAA Privacy Rule. The HIPAA-driven penalties associated with privacy breaches, the HITECH Act expansion of penalties to business associates, state laws regarding privacy and confidentiality, and lack of perceived incentives for linkage...
are significant. Because HIPAA impedes linkage of EMR with claims data, it also impedes patient-centered outcomes research and quality-of-care evaluations, among other research endeavors.

To address the anticipated need to link data via identifiers, the IOM has recommended establishment of “certified entities” that link and match personally identifiable information from multiple organizations so that longitudinal data sets are created that remove direct identifiers. Because of the current and probably future lack of a unique health-care identifier to link the data accurately, this valuable step forward would best take advantage of the increasing availability of electronic health data.

Acknowledgements

The authors appreciate the assistance of Beth Lasater, Kristen Rosati, and Tameka Sama with obtaining institutional review board approvals at RTI International and Arizona State University and with executing the business associate agreements between Arizona State University and the commercial EMR vendor.

Funding

The research presented in this article was funded by RTI International.

References

Appendix 1

The Health Insurance Portability and Accountability Act and protected health information

The Health Insurance Portability and Accountability Act (HIPAA) was passed by the US Congress in 1996. The primary goals of this legislation were to improve people’s ability to continue their health insurance, regardless of their employment status; reduce fraud; and simplify the administration of health coverage. The HIPAA Privacy Rule, developed under the administrative simplification provisions, required the adoption of standards to protect individually identifiable health information. This information is called protected health information (PHI) under HIPAA. Information is considered PHI if it contains any of the following 18 specific identifiers, is related to a person’s health or treatment, and is held or transmitted by a covered entity:

- Name;
- Telephone numbers;
- Any elements of dates (other than year);
- Geographic subdivisions smaller than a state (i.e. street address, city, county, precinct, zip code, and geocodes, except for the initial three digits of a zip code);
- Fax numbers;
- Electronic mail addresses;
- Social Security numbers;
- Medical record numbers;
- Health plan beneficiary numbers;
- Account numbers;
- Certificate/license numbers;
- Vehicle identifiers and serial numbers;
- Device identifiers and serial numbers;
- Web universal resource locators;
- Internet protocol address numbers;
- Biometric identifiers (finger and voice prints);
- Full-face photographic images and any comparable images;
- Any other unique identifying number.

Covered entities, as defined by HIPAA, include health-care providers, health plans, and health-care clearinghouses. The Privacy Rule states, that covered entity may not use or disclose protected health information, except as permitted or required by this legislation.

A covered entity may contract with a business associate under a business associate agreement (BAA) to help with activities related to treatment, payment, or operations. A BAA is required only when a person or entity is conducting a function or activity regulated by the Administrative Simplification Rules on behalf of a covered entity, such as providing one of the services listed in the definition of “business associate” (see 45 C.F.R. § 160.103). Under this provision, BAAs help covered entities with activities covered by HIPAA, but they are not intended for research activities.
The use or disclosure of PHI for health research is permitted by the Privacy Rule under its research provisions. When obtaining PHI from a covered entity for research, a researcher must either (1) obtain individual authorization (permission) from each patient to use his or her health information (45 C.F.R. § 164.508) or (2) use one of the research provisions that do not require authorization (45 C.F.R. § 164.512). The requirements that, when met, allow release of PHI to researchers without patients’ authorization are as follows:

- Obtaining a waiver of authorization from an institutional review board (IRB) or privacy board;
- Receiving a “limited data set” from which direct patient identifiers have been removed;
- Using the PHI only for activities “preparatory to research” (e.g. determining whether there are sufficient numbers of patients with a certain condition to do a valid study);
- Using only PHI about deceased individuals.

The Privacy Rule defines each of these options and under what circumstances each can be used. Obtaining authorization from each patient whose data are to be linked is impracticable for linkage research because, typically, hundreds of thousands of patient records are used. The second option (limited data set) is unfeasible because linkage requires direct identifiers, and the last two options (preparatory to research and decedent research) are not applicable to linkage research. Therefore, the waiver of authorization is the only HIPAA research option available to linkage researchers. The waiver of authorization option requires that the use or disclosure of the PHI pose “no more than minimal risk to the privacy of the individual.” This requirement is very hard to meet if the link between the databases is a direct identifier (or combination of identifiers), especially if one of the identifiers is the Social Security number. If the researchers cannot convince the IRB or privacy board that this requirement is met—by masking the identifiers or using a trusted third party to do the linkage—then the board is unlikely to grant the waiver. Still, the HIPAA waiver of authorization is likely the best option for data linkage research.