Does Access Modality Matter? Evaluation of Validity in Reusing Clinical Care Data

Christopher P. Danford[1], Monica M. Horvath, PhD[2], W. Edward Hammond, PhD[2], Jeffrey M. Ferranti, MD, MS[2,3]

[1]Duke University School of Medicine, Durham, NC; [2]Duke Health Technology Solutions, Duke University Health System, Durham, NC; [3]Department of Pediatrics, Duke University School of Medicine, Durham, NC

Abstract

Self-service database portals may improve access to institutional data resources for clinical research or quality improvement, but questions remain about the validity of this approach. We tested the accuracy of data extracted from a clinical data repository using a self-service portal by comparing three approaches to measuring medication use among patients with coronary disease: (1) automated extraction using a portal, (2) extraction by an experienced data architect, and (3) manual chart abstraction. Outcomes included medications and diagnoses (e.g., myocardial infarction, heart failure). Charts were manually reviewed for 200 patients. Using matched criteria, self-service query identified 7327 of 7358 patients identified by the data analyst. For patients in both cohorts, agreement rates ranged from 0.99 for demographic data to 0.94 for laboratory data. Based on chart review, the self-service portal and the analyst had similar sensitivities and specificities for comorbid diagnoses and statin use.

Background

Over the last decade, electronic health record systems and their associated data repositories at large medical centers have evolved to meet institutional requirements for public reporting and quality improvement (QI). The secondary use of data captured electronically during routine clinical care has produced substantial benefits in efficiency[1, 2]. Furthermore, strong interest exists to capitalize on clinical data repositories for translational research and trial recruitment. As these data resources have matured, demand for access from clinicians has increased dramatically and in many cases has outstripped the availability of data analysts to fill requests. Self-service access has emerged as a solution to this bottleneck: according to a 2010 survey of Clinical and Translational Science Award informatics directors, the number of organizations providing self-service access to their integrated data repository (IDR) doubled between 2008 and 2010, and providing self-service access was a top priority for future enhancements[3].

Previously, for data extraction from the IDR at our institution, clinicians with research questions or QI personnel submitted data requests to analysts with data modeling expertise. These technical experts wrote custom structured query language (SQL) queries according to clinical specifications for each request. Recently, the Duke University Health System developed a self-service research portal, the Duke Enterprise Data Unified Content Explorer (DEDUCE), that enables clinicians and QI personnel to query the IDR directly with institutional review board approval[4].

This self-service tool and others like it add a layer to data extraction that has yet to be validated. Although self-service portals hold promise, secondary use of clinical data for research, reporting, or QI is predicated on the data corresponding to an accepted reference standard—in other words, their accuracy.

To assess the validity of our self-service portal, DEDUCE, we defined a cohort using the portal and matched them to a cohort defined by custom SQL query; target outcomes were then extracted from the data repository using DEDUCE and SQL query, and these were compared to manual chart review. We hypothesized that if use of a self-service portal were equivalent to SQL query by a data analyst, data exported from DEDUCE would be identical to data exported by SQL query, and both sets of data would reflect data abstracted from manual chart reviews.

Methods

Data sources

The data source for electronic extraction was the Duke Medicine enterprise data warehouse (EDW), which contains the records of all 4 million Duke University Health System patients, with more than 20 million encounters per year from three hospitals and 116 ambulatory clinics. The EDW is a dimensionally modeled, standards-based store of structured data organized into high-level subject areas managed by a central data governance committee. Available data include demographics, encounters, clinical orders, lab results, medications, diagnoses, and procedures. A formal extract, transform, and load procedure integrates data from source systems on a nightly basis to ensure consistency and quality and to minimize redundancy.

Access to the EDW is possible through SQL query or the DEDUCE portal. Analysts execute SQL queries using a variety of commercial and freeware SQL clients, as well as the SAS software package (SAS Institute, Inc., Cary, NC). DEDUCE directly accesses the EDW, and cohorts are built using a web-based ASP.NET and C# application that provides a graphical interface for querying a subset of subject areas. Details and screenshots of this portal have been published previously[4].

Study population

The content area for this study was lipid management in an ambulatory setting by cardiology specialists. We included in the analysis any patient 18 years of age or older with coronary artery disease who had at least two visits with a Duke Cardiology provider during the 13-month period from 6/1/2009 to 6/30/2010. We defined patients with coronary artery disease using the *International Classification of Diseases, Ninth Edition, Clinical Modification* (*ICD-9-CM*) codes 410-412 (any suffix), 414 (any suffix), V45.81, or V45.82 coded at any encounter prior to the start of the study period or in the first 6 months of the study period (through 12/1/2009). Outcome variables included date of birth, gender, race, low-density lipoprotein cholesterol (LDL-C) measurement, statin use, and comorbid diagnoses of previous myocardial infarction (MI), hyperlipidemia, diabetes mellitus, heart failure, chronic kidney disease, end-stage renal disease, cerebrovascular disease, and liver disease. Comorbid diagnoses were defined by the *ICD-9-CM* codes listed in Table 1.

Table 1. Comorbid diagnoses by *ICD-9-CM* classification.

Comorbidity	***ICD-9-CM*** Code
Previous Myocardial Infarction	410.xx
Hyperlipidemia	272.xx
Diabetes Mellitus	250.xx, V58.67
Heart Failure	398.91, 402.x1, 404.x1, 404.x3, 428.xx
Chronic Kidney Disease	403.xx, 404.xx, 585.xx
End-Stage Renal Disease	585.6
Cerebrovascular Disease	430.xx-438.xx
Liver Disease	570.xx-573.xx

Self-service portal versus custom SQL query

Patient cohorts were identified and target outcomes extracted from the EDW in two ways: by study personnel using DEDUCE and by a data analyst using SQL code. All DEDUCE queries were executed by one of the clinical authors, who had undergone 3 hours of training required of all new users of the data portal, as well as about 1 year of experience using DEDUCE. Data elements extracted were EDW-specific patient identifiers; date of ambulatory cardiology encounters; dates and results of LDL-C, high-density lipoprotein cholesterol (HDL-C), and total cholesterol measurements; medication data associated with all encounters during the study period; and *ICD-9-CM* diagnoses corresponding to targeted comorbidities. The SQL queries were executed by a data analyst with over 10 years of experience with DUHS data resources. Inclusion criteria for the patient cohorts were developed in collaboration between the authors and data analyst before all queries were made.

We used patient identifiers to compare outcome variables between datasets on a patient-by-patient level. For each patient in each dataset, statin use was a binary variable defined as positive if a statin medication was associated with any ambulatory encounter during the study period. Each comorbid condition was a binary variable defined as positive if the patient had ever had an encounter coded with one of the *ICD-9-CM* codes listed in Table 1. LDL-C, HDL-C, and total cholesterol were defined by the lowest value of each laboratory test if any value was available. A second variable was defined as positive if the patient's lowest LDL-C measurement during the study period was < 100 mg/dL.

Manual chart abstraction

To evaluate provider handling of lipids in the outpatient setting, charts were reviewed for a randomly selected, stratified subset of patients defined by the custom SQL query. A total of 200 patients were randomly selected for review: 100 patients identified in the SQL dataset as not taking a statin, 50 patients taking a statin with a documented statin allergy, and 50 patients taking a statin with no statin allergy. All charts were reviewed by the first author, who abstracted cardiology notes during the study period, documenting comorbidities and statin prescription based on the problem list, medication list, and plan. All variables were defined using a data dictionary developed before review. For three of the comorbid diagnosis variables—diabetes mellitus, chronic kidney disease, and hyperlipidemia—further review was performed, and variables were classified as positive if patients met laboratory-defined criteria for disease (hemoglobin A1c > 6.5%, estimated glomerular filtration rate < 60 mL/min/1.73 m^2, or LDL-C > 100 mg/dL, respectively) regardless of whether the condition was documented in a provider note.

The first author was blinded to the results of DEDUCE and SQL queries at the time of chart abstraction. Charts were reviewed using a web application that was the primary clinical information system for providers at the time of this study. This application provided views into a repository of discharge summaries, physician documentation, orders, allergies, procedure histories, laboratory results, and radiology reports as well as scanned notes from other information systems not yet integrated.

Statistical analysis

For patients with coronary artery disease identified in both cohorts, agreement between DEDUCE and SQL query were measured using simple proportion agreement and Cohen's kappa for binary variables to account for chance agreement. Two-by-two tables were constructed to assess agreement between both SQL and DEDUCE data and manual chart review for the subset of 200 charts. All statistical analyses were carried out using JMP 10.0 (SAS Institute, Inc., Cary, NC). Ethical approval was obtained from the Duke University Institutional Review Board (PRO27093).

Results

Sample description

The DEDUCE query portal defined a cohort of 7327 patients who met the study criteria. SQL query identified 7324 of the 7327 individuals identified by the DEDUCE query portal as well as 34 additional patients. The additional patients identified by SQL query were all attributed to a provider who has a dual appointment in Cardiology and Pulmonary-Critical Care (see Discussion below). Patient characteristics of both cohorts are defined in Table 2.

Table 2. Characteristics of cohort.

Characteristic	DEDUCE Access	SQL Access
Mean Age, y	67.2 (95% CI, 66.9-67.5)	66.9 (95% CI, 66.6-67.2)
Female, n	2637 (36%; 95% CI, 35%-37%)	2656 (36%; 95% CI, 35%-37%)
Race, n		
White	5522 (75%; 95% CI, 74%-76%)	5541 (75%; 95% CI, 74%-76%)
Black	1388 (19%; 95% CI, 18%-20%)	1398 (19%; 95% CI, 18%-20%)
American Indian	236 (3.2%; 95% CI, 2.8%-3.7%)	236 (3.2%; 95% CI, 2.8%-3.7%)
Asian	75 (1.0%; 95% CI, 0.8%-1.3%)	77 (1.1%; 95% CI, 0.8%-1.3%)
Other	106 (1.5%; 95% CI, 1.2%-1.8%)	106 (1.4%; 95% CI, 1.2%-1.8%)
Mean LDL-C, mg/dL	81.5 (95% CI, 80.8-82.2)	85.3 (95% CI, 84.5-86.1)
Total, n	7327	7358

Agreement between DEDUCE and SQL query

For patients identified in both cohorts, the proportion of agreement between DEDUCE and SQL query was highest for gender, race, and date of birth (Table 3). Rates of agreement for all variables were 0.94 or greater, with lowest agreement for date and value of the lowest LDL-C measurement. Rates of agreement on whether the patient was at goal with respect to LDL-C and had received statin therapy during the study period were 0.99 and 0.98, respectively.

Table 3. Agreement between DEDUCE and SQL query.

Outcome Variable	Kappa	Proportion Agreement
Date of Birth		0.999
Gender		0.999
Race		0.999
Previous Myocardial Infarction	0.97	0.99
Hyperlipidemia	0.88	0.97
Diabetes Mellitus	0.96	0.98
Heart Failure	0.94	0.97
Chronic Kidney Disease	0.91	0.97
Cerebrovascular Disease	0.94	0.98
End-Stage Renal Disease	0.93	0.99
Liver Disease	0.94	0.99
Low LDL-C		0.94
Low HDL-C		0.94
Low Total Cholesterol		0.94
Statin Use	0.93	0.98
LDL-C < 100 mg/dL	0.98	0.99

Kappa was calculated for agreement between binary variables only.

Comparison with chart abstraction

A total of 254 charts were reviewed. If the patient did not have coronary artery disease by chart review, the chart was not reviewed further, with the result that 200 charts underwent structured abstraction. DEDUCE and SQL query had similar sensitivity and specificity with respect to chart review (Table 4). Classification accuracy was greatest for statin prescription, diabetes, and end-stage renal disease. Ability to detect disease with *ICD-9-CM* codes either by DEDUCE or by the data analyst was lowest for previous MI, chronic kidney disease, and liver disease, with sensitivities of 0.59, 0.59, and 0.50, respectively. Specificities were lowest for hyperlipidemia and heart failure with 0.53 and 0.58, respectively.

Table 4. Sensitivity and specificity of different access modes.

	DEDUCE			SQL Query		
Variable	Sensitivity (95%CI)	Specificity (95%CI)	Kappa	Sensitivity (95%CI)	Specificity (95%CI)	Kappa
Previous Myocardial Infarction	0.59 (0.46-0.71)	0.87(0.80-0.92)	0.47	0.62 (0.49-0.74)	0.87 (0.80-0.92)	0.50
Hyperlipidemia	0.93 (0.88-0.96)	0.53 (0.29-0.75)	0.42	0.95 (0.90-0.98)	0.47 (0.25-0.71)	0.43
Diabetes Mellitus	0.86 (0.77-0.91)	0.94 (0.86-0.97)	0.79	0.89 (0.81-0.94)	0.94 (0.86-0.97)	0.83
Heart Failure	0.88 (0.78-0.94)	0.58 (0.48-0.66)	0.42	0.91 (0.82-0.96)	0.55 (0.46-0.64)	0.41
Chronic Kidney Disease	0.59 (0.48-0.69)	0.93 (0.87-0.97)	0.52	0.60 (0.50-0.70)	0.90 (0.82-0.95)	0.51
Cerebrovascular Disease	0.74 (0.55-0.87)	0.73 (0.65-0.79)	0.31	0.74 (0.55-0.87)	0.71 (0.63-0.78)	0.29
End-Stage Renal Disease	1.00 (0.52-1.00)	0.98 (0.94-0.99)	0.79	1.00 (0.52-1.00)	0.98 (0.94-0.99)	0.75
Liver Disease	0.50 (0.26-0.74)	0.97 (0.93-0.99)	0.50	0.50 (0.26-0.74)	0.96 (0.92-0.98)	0.46
Statin Use	0.98 (0.92-1.00)	1.00 (0.95-1.00)	0.98	0.87 (0.79-0.93)	1.00 (0.95-1.00)	0.85

Discussion

The purpose of this study was to assess the validity of a self-service portal for secondary use of electronic health data. We hypothesized that if our self-service data portal was equivalent to current methods of data extraction, then data exported from the EDW with DEDUCE would match data exported by a data analyst using SQL query, and both would have similar sensitivity and specificity with respect to manual chart review. Furthermore, we suspected that certain claims-based definitions of disease were more accurate than others.

Self-service portal versus custom SQL query

Comparisons of data extracted from the EDW using DEDUCE versus SQL query showed a high rate of agreement at a patient level. Static data elements such as date of birth, gender, and race were most similar between the two methods of data access. Comorbid diagnoses were defined by exporting *ICD-9-CM* codes using both access methods and then classifying patients in a binary way. Although levels of agreement were lower than for demographic data, these results showed high levels of agreement. The lowest levels of agreement were for individual LDL-C measurements, but even these variables achieved greater than 90% agreement. Overall, these results suggest that static variables are most accurate, individual data elements have greater than 90% accuracy, and composite outcome variables improve their accuracy.

Differences between the self-service and SQL datasets could have several different causes. Although we sought to match selection criteria used in the DEDUCE and SQL queries as much as possible, interpretation of criteria is subjective, and search strategies could differ. As an example, DEDUCE used a table listing "senior provider specialty" to define visits with a cardiologist, while the data analyst used a local fact table of cardiology providers that the analyst could reference in constructing SQL queries. This table included one provider with a dual appointment in Cardiology and Pulmonary-Critical Care medicine, resulting in inclusion of a small number of extra patients in the SQL cohort. Other discrepancies between DEDUCE and SQL query results could be because DEDUCE queries a subset of the EDW; in certain cases, such as for patients with no entered date of birth, the SQL query could access those patients while DEDUCE could not. Lastly, the SQL and DEDUCE queries were made at different times, and so some discrepancies between the SQL and DEDUCE queries could be due to updates to the EDW between each query.

Self-service portal and SQL query versus manual chart review

The DEDUCE and SQL queries had similar sensitivity and specificity with respect to manual chart review. Among the comorbidities that we assessed, classification accuracy of our *ICD-9-CM* map varied by condition; several had both high sensitivity and specificity (e.g., diabetes and end-stage renal disease). On the other hand, some of our *ICD-9-CM* definitions of comorbidity had high sensitivity but low specificity, meaning that a history of *ICD-9-CM* code was good for detecting cases of the disease but also included many individuals who did not truly have the condition. Low specificity is consistent with inclusion of too many individual ICD codes in our definition, caused by coding conditions for an encounter when in fact they were ruled out or chronic conditions that wax and wane (e.g., heart failure in a cohort of patients with known coronary artery disease).

Other claims-based definitions of disease had high specificity but low sensitivity, reflecting poorer case detection but better distinction between healthy and sick individuals. Low sensitivity could be due to an overly inclusive definition of disease on chart review, an overly narrow selection of *ICD-9-CM* codes for definition of the disease, or a condition that was often noted in the problem list or plan but not coded (e.g., history of MI). In our case, our definition of liver disease on chart review included patients with a history of elevated transaminases, although our *ICD-9-CM* definition of disease did not include code 790.4, "Nonspecific elevation of levels of transaminase or lactic acid dehydrogenase [LDH]." Also, our chart review definition of chronic kidney disease included all patients who had had an estimated glomerular filtration rate of <60 mL/min/1.73 m^2 for more than 3 months not in the setting of an acute rise in creatinine. Based on the low sensitivity of this diagnosis, it would appear that physicians do not always recognize this as chronic kidney disease or at least do not code it as such.

The differences between the manual chart review and electronic data highlight the fact that many current clinical workflows are not optimized to collect electronic health data in a way that can be reused for QI or research. Consequently, data are often not entered in coded format by the clinician, clinicians sometimes fail to use the information systems that feed the clinical data repository, coding errors arise from misunderstandings of codes by clinicians or coders, or data are available to the clinician that are not part of the clinical data repository.

Defining disease states with administrative data

Regardless of access modality, it remains difficult to judge the classification accuracy of claims-based definitions of disease on natural populations of patients. Without a reference set, manual review is the only way to distinguish cases from non-cases, which limits accurate estimation of specificity, and there is no way to define true positives a priori, which limits estimation of sensitivity. Resources are needed to better define standard electronic definitions of disease. Recently, the emergence of IDRs has made this research much easier, but problems remain, and these can often be subtle[5, 6]. Also, text analytic methods are on the horizon, which could obviate the need for coded data, but these methods face the challenge of balancing between accuracy and computation times when scaled up for enterprise-wide data warehouses[7, 8].

Limitations

Limitations of this study include a small sample size for chart review and the use of a single reviewer. Errors in coding arise when codes are applied incorrectly, conditions are suspected initially and then ruled out, or when some codes represent etiologies of disease rather than sequelae. Also, the results of this study are system-specific and not generalizable to other systems. Lastly, our study used only a single SQL query and a single DEDUCE query.

Conclusion

Measuring the accuracy of electronically captured data is a crucial first step in secondary use of the data for QI or clinical research. We know of no other studies to date comparing self-service data access by clinicians versus SQL query by data analysts. Our findings suggest that self-service portals provide a valid form of access to IDRs. Data extracted using the DEDUCE portal showed high agreement with data extracted by SQL query, and datasets produced using both access methods showed similar sensitivity and specificity with respect to manual chart review. Our findings also confirm that difficulties surround the use of *ICD-9-CM* codes for definition of disease states, regardless of the means used to access an IDR. Further research is needed to define standards for defining disease states using clinical variables and linked encounters available in IDRs.

References

1. Tang PC, Hammond WE. A progress report on computer-based patient records in the United States. In: Dick RS, Steen EB, Detmer DE, eds. The Computer-Based Patient Record: An Essential Technology for Healthcare. 2nd ed. Washington DC: National Academy Press; 1997:1-20.
2. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc 1997;4:342-355.
3. Mackenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc 2012;19:e119-e124.
4. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement. J Biomed Inform 2011;44:266-276.
5. Iezzoni LI, Ash AS, Shwartz M, Landon BE, Mackiernan YD. Predicting in-hospital deaths from coronary artery bypass graft surgery: do different severity measures give different predictions? Med Care 1998;36:28-39.
6. Lindenauer PK, Lagu T, Shieh MS, Pekow PS, Rothberg MB. Association of diagnostic coding with trends in hospitalizations and mortality of patients with pneumonia, 2003-2009. JAMA 2012;307:1405-1413.
7. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;306:848-855.
8. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15:14-24.