Journal of the American Medical Informatics Association, 24(6), 2017, 1134–1141
doi: 10.1093/jamia/ocx071
Advance Access Publication Date: 1 August 2017
Research and Applications

Research and Applications

Biases introduced by filtering electronic health records for patients with "complete data"

Griffin M Weber,[1,2] William G Adams,[3] Elmer V Bernstam,[4] Jonathan P Bickel,[5] Kathe P Fox,[6] Keith Marsolo,[7] Vijay A Raghavan,[8] Alexander Turchin,[9] Xiaobo Zhou,[10] Shawn N Murphy,[11] and Kenneth D Mandl[1,5]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA, [2]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA, [3]Department of Pediatrics, Boston Medical Center, Boston, MA, USA, [4]Department of Internal Medicine, McGovern Medical School, School of Biomedical Informatics, The University of Texas Health Science Center, Houston, TX, USA, [5]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA, [6]Department of Analytics and Behavior Change, Aetna, Hartford, CT, USA, [7]Department of Pediatrics, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, OH, USA, [8]Scientific Information Management, Merck, Boston, MA, USA, [9]Division of Endocrinology, Brigham and Women's Hospital, Boston, MA, USA, [10]Department of Radiology, Wake Forest University School of Medicine, Winston Salem, NC, USA and [11]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

Corresponding Author: Griffin M Weber, 10 Shattuck St, Boston, MA 02115, USA. E-mail: weber@hms.harvard.edu. Phone: 617-432-6134

Received 14 January 2017; Revised 3 June 2017; Accepted 12 June 2017

ABSTRACT

Objective: One promise of nationwide adoption of electronic health records (EHRs) is the availability of data for large-scale clinical research studies. However, because the same patient could be treated at multiple health care institutions, data from only a single site might not contain the complete medical history for that patient, meaning that critical events could be missing. In this study, we evaluate how simple heuristic checks for data "completeness" affect the number of patients in the resulting cohort and introduce potential biases.

Materials and Methods: We began with a set of 16 filters that check for the presence of demographics, laboratory tests, and other types of data, and then systematically applied all 2^{16} possible combinations of these filters to the EHR data for 12 million patients at 7 health care systems and a separate payor claims database of 7 million members.

Results: EHR data showed considerable variability in data completeness across sites and high correlation between data types. For example, the fraction of patients with diagnoses increased from 35.0% in all patients to 90.9% in those with at least 1 medication. An unrelated claims dataset independently showed that most filters select members who are older and more likely female and can eliminate large portions of the population whose data are actually complete.

Discussion and Conclusion: As investigators design studies, they need to balance their confidence in the completeness of the data with the effects of placing requirements on the data on the resulting patient cohort.

Key words: electronic health records, claims data, data accuracy, information storage and retrieval, selection bias

INTRODUCTION

A health care system instrumented to learn and discover will derive knowledge and drive progress with data collected during the routine care of patients. These data will fuel myriad efforts, such as quality improvement activities, surveillance, observational cohort studies, pragmatic clinical trials, and disease association studies. The wide-scale adoption of electronic health records (EHRs) over the past 6 years has dramatically increased the availability of electronic clinical data.

A frequently cited challenge for the secondary use of these clinical data has been the lack of standards on how health care information is collected and stored. But an equally important issue is data completeness.[1–3] To streamline clinical trials and embed them in the health care system (pragmatic trials), or to conduct trials entirely in silico (observational studies), the data need to be complete, allowing patient outcomes to be tracked efficiently and effectively simply by using datasets produced as a byproduct of care, as opposed to using prospective data collection.

An observational trial comparing the use of dabigatran vs warfarin, for example, will need an accounting of the number of complications – thrombotic events and bleeds – in each study arm. But if the EHR data are derived from only a subset of health care institutions, they will not provide a complete view of a patient seeking care across sites.[4–9] Though a patient may have her cardiology outpatient visits in health system A, she might be brought to the emergency department of health system B if she has a stroke.

One approach to addressing this problem is to concatenate multiple data sources to piece together a patient's record across sites of care. For clinical care delivery, health information exchanges have attempted this, with varying degrees of success.[10–13] To drive accountable care, risk-bearing organizations are beginning to merge their local EHR data with payor data sources. The payor data sources tend to have a complete record of all services billed for during the insurance enrollment period, regardless of site of care.[14] However, substantial issues impede the linkage of claims data to EHR data, such as regulatory oversight, privacy, data rights, and the lack of a universal patient identifier.

As an alternative to merging data across institutions, investigators can search for patients within each site whose data are "complete enough." What that means depends on the particular characteristics of the study. For example, a study might require the diagnoses and medications of its patient cohort but not need any laboratory test results. Simple heuristic filters that select only patients with 1 or more types of data are easy to implement and sufficient for many research questions. However, to our knowledge, this approach has never been evaluated on a large scale at multiple institutions. In this study, we determine the effect of different combinations of data completeness filters on the number of patients in the resulting cohort and examine the potential biases that the filters introduce.

MATERIALS AND METHODS

Figure 1 illustrates the overall workflow of this study. The primary analysis is based on EHR data, and a secondary analysis was conducted using an unrelated administrative claims dataset. The results of running the data completeness filters on the 2 types of data were qualitatively compared. (Note that a direct comparison between EHR and claims data was not possible, because the data are not linked at the patient level and include partially overlapping but mostly different patient populations.)

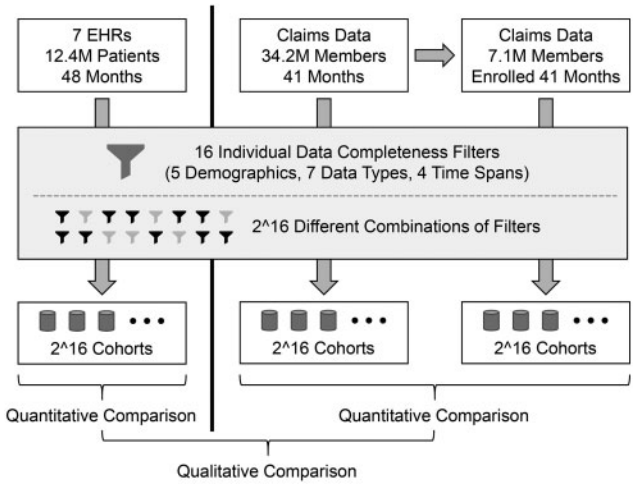

Figure 1. Overall project structure and analysis workflow. Three sets of experiments were conducted: first on EHR data for 12.4 million patients, second on claims data for a different group of 34.2 million members, and third on the subset of 7.1 million members in the claims dataset who were continually enrolled for all 41 months. Because the EHR and claims data represent different populations, the results of those separate experiments can only be compared qualitatively. (The results of the second experiment are only presented in the supplementary material.)

EHR data sources

The setting was 7 hospitals and health systems in a clinical data research network designed for pragmatic and observational trials, the Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS).[15] SCILHS is part of PCORNet, a national network funded by the Patient-Centered Outcomes Research Institute under the Affordable Care Act.[16,17] Each SCILHS site maintains a repository of EHR data that are generated as byproducts of clinical care delivery. While the information in EHRs is site-specific, EHRs provide granular clinical data, textual notes, and laboratory results. The number of patients in each repository ranges from 655 902 to 3 010 950, with a total of approximately 12 million patients in the network (because patient records are not linked across EHRs, this is likely an overestimate of the actual number of distinct individuals). The data from each site span a 4-year period, from July 1, 2010, through June 30, 2014. The SCILHS sites are: Beth Israel Deaconess Medical Center, Boston Children's Hospital, Boston Medical Center, and Partners Healthcare (which includes Brigham and Women's Hospital and Massachusetts General Hospital) in Boston, Massachusetts; Cincinnati Children's Hospital Medical Center in Cincinnati, Ohio; The University of Texas Health Science Center in Houston, Texas; and Wake Forest Baptist Medical Center in Winston-Salem, North Carolina. All 7 SCILHS sites obtained Institutional Review Board approval to conduct this study.

Heuristic computational filters

A simple approach to finding subsets of patients whose data are suitable for research studies is to use heuristic computational filters that exclude patients who are missing different types of data from their records. The challenge in using these filters is distinguishing patients with missing data from patients who are relatively healthy or who have not sought medical care recently. Each of these types of patients will have a low number of data facts in their records. As a result, these filters might bias the resulting cohort, selecting sicker patients who interact with the health care system more often. For example, in a cohort of 10 000 patients, Rusanov previously found

that patients with worse health status (as defined by the American Society of Anesthesiologists Physical Status Classification System) had more laboratory test results and medication orders.[18] However, because these simple data completeness checks are easy to implement, and an alternative method might not be available, they are the best option for many studies.

The starting point for the filters we consider in this study was 2 previous internal projects we conducted within 1 of our hospitals: 1 project to search the EHR for patients who receive primary care at that site, and another to search for healthy patients. Based on our experience from those projects and knowledge about the types of studies investigators are planning to run on our SCILHS network, we evaluated 16 filters (Table 1), divided into 3 groups:

(1) The first group of filters is based on patient demographics. Two filters check whether the patients have a recorded race (Race) and both age and sex (AgeSex). The Alive filter excludes patients who are known to be deceased. The SameState filter selects patients who live in the same state as the health care system, since out-of-state patients may be more likely to receive additional care closer to their homes. At hospitals that treat all age groups, the AgeCutoffs filter selects all patients. At pediatric hospitals (2 SCILHS sites), the AgeCutoffs filters selects only patients who are currently <30 years old, since older patients are likely treated at adult care facilities. Similarly, at hospitals that primarily see adult patients (1 site), the AgeCutoffs filter selects only patients ≥20 years old.

(2) Data fact type filters check whether patients have at least 1 recorded diagnosis (Diagnoses), vital sign (VitalSigns), laboratory test result (LabTests), medication (Medications), or outpatient visit (OutpatientVisit). The RoutineVisit filter selects 3 specific International Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes frequently associated with primary care visits: V20.2 (routine infant or child health check), V70.0 (routine general medical examination at a health care facility), and V72.31 (routine gynecological examination). We defined "fact count" as the total number of diagnoses, vital signs, laboratory test results, medications, and procedure codes each patient has. The NoSmallFactCount filter excludes patients who have no data facts. Among patients with data facts, the NoSmallFactCount filter also excludes patients who are in the bottom 10% of patients in terms of fact count.

(3) Time span filters select for patients with multiple interactions with the health care system. The overall time window for this study was the 4-year period from July 1, 2010, through June 30, 2014. The FirstLast18Months and FirstLastYear filters select patients with at least 1 data fact in both the first and last 18 months and both the first and last year, respectively. The All1YearPeriods and All6-MonthPeriods filters require at least 1 fact in each of the four 1-year or eight 6-month time blocks, respectively.

These 16 filters can be mapped to 4 metrics of EHR completeness, as defined by Weiskopf[19]: "Breadth" describes the types of available data, which corresponds to the AgeSex, Race, NoSmall-FactCount, Diagnoses, VitalSigns, LabTests, and Medications filters. "Density" is the number and frequency of data points over time, which corresponds to the FirstLast18Months, FirstLastYear, All1YearPeriods, and All6MonthPeriods filters. "Documentation" indicates whether all observations about a patient were recorded. This is difficult to measure. As a proxy, Weiskopf looked for the presence of clinical notes, since each visit should have a note. The AgeCutoffs, SameState, OutpatientVisit, and RoutineVisit filters are a different proxy that looks for evidence of primary care visits, since most patients with complete data should have those. Finally, "Predictive" refers to the availability of study-specific data elements needed to predict a particular outcome. The Alive filter addresses the SCILHS use case of predicting which patients are eligible for a clinical trial (ie, they must be alive to be enrolled).

Table 1. Simple heuristic filters for selecting patients with relatively complete data

Filter	Description
Demographics	
AgeSex	Has both age and sex
AgeCutoffs	Age <30 years if few adults, ≥20 years if few children
Race	Has race
Alive	Is alive
SameState	Lives in same state as health care center
Data Fact Types	
NoSmallFactCount	Is not in the bottom 10% of total data fact count
Diagnoses	Has diagnoses
VitalSigns	Has vital signs
LabTests	Has laboratory tests
Medications	Has medications
OutpatientVisit	Has outpatient visits
RoutineVisit	Has a routine visit ICD-9 code (V70.0, V72.31, V20.2)
Time Spans	
FirstLast18Months	Has data in the first and last 18-month blocks
FirstLastYear	Has data in the first and last year
All1YearPeriods	Has data in all 1-year blocks
All6MonthPeriods	Has data in all 6-month blocks

Additional details about the choice of filters and how they are computed are provided in the supplementary material. Note that the purpose of this study was simply to estimate the biases caused by different types of filters that researchers might apply to EHR to check for data completeness. We do not claim that these are the most common filters used by researchers or that they are the best filters for selecting patients with complete data.

Claims data

While an individual might receive care at multiple health care systems, individuals enroll in a given insurance plan for defined and often extended periods of time. Payor claims are produced to bill for provided health services. Since all the health care interactions of beneficiaries are captured, claims data permit a holistic view of patients while they are enrolled in a health plan. This provides an opportunity to evaluate the biases introduced by the heuristic filters by applying them to a group of patients whose data are known to be complete over a given period of time. For this study, we used a nationwide payor claims database provided by Aetna health insurance with data on 34 184 719 members over a 41-month period from January 2010 through May 2013. From this, we selected the 7 099 393 members who were continuously enrolled for the entire 41 months. The main results below are based solely on the 7.1 million continuously enrolled members. The supplementary material describes additional analyses performed on the entire 34.2-million-member dataset. (Note that although the claims datasets include large geographically and demographically diverse populations, they do not equally represent all people in the United States. Also, they do not exclude individuals who had other forms of health coverage at the same time as they had Aetna insurance.)

Implementation details

Each SCILHS site stores its EHR data in an open-source clinical data repository called Informatics for Integrating Biology and the Bedside (i2b2).[20] A database script was implemented to determine which patients in an i2b2 database passed different combinations of filters. For each set of patients, the script returned the number of patients; the mean fact count per patient; the fraction of patients with diagnoses, routine visit ICD-9 codes, outpatient visits, inpatient visits, and emergency room visits; and breakdowns by sex and age group. The age group was defined as floor (age/10)+5, so that, for example, all patients between 50 and 59 years old were in age group 55. Because SCILHS sites exchanged age groups rather than age in years, we could only calculate "mean age group" instead of mean age. The database script is available in the "Loyalty Cohort" project on https://community.i2b2.org.

The claims data were also loaded into an i2b2 repository, enabling us to run the same database script that was applied to the EHR data. However, because the claims data did not include race, vital signs, visit type or location, or whether members were still alive, it was not possible to run all filters. Also, because the claims data represented only 41 months, the actual fact counts from the data were multiplied by 48/41 to estimate the fact counts over a 48-month period. These 48-month estimates are the fact counts reported in this study for the claims data.

RESULTS

Filters applied to EHR data

The 16 filters were first individually applied to the EHR data to determine their separate effects. Although there was considerable variation across health care systems, in general, most patients had demographic data and passed the group 1 filters, about one-quarter had any given data fact type and passed the group 2 filters, and about 10% had multiple facts and passed the group 3 filters (Figure 2).

To estimate bias introduced by the filters, Table 2 lists aggregate statistics for the patients who passed each of the filters. Without any filters applied, the patient population as a whole had a mean of 63.2 data facts over the 4-year period. Since the fact count combines all types of health data, we will use it here as a rough approximation for overall patient health status. With the exception of the Alive filter, all other filters resulted in cohorts with mean patient fact counts >63.2. The time span filters produced cohorts with the highest mean fact counts, the largest of which is the All6MonthPeriods filter, which selected patients with a mean of 787.7 facts.

Table 2 also shows that all the filters selected cohorts whose patients, when compared to the entire hospital patient population, had a higher mean number of diagnoses and routine, outpatient, inpatient, and emergency room visits. Filters that selected patients with higher mean fact counts also tended to have a larger fraction of female patients.

The 16 filters can be combined in $2^{16} = 65\,536$ ways, including the case where no filters are applied and all patients are selected. Figure 3 illustrates the effect of each of these combinations on the resulting cohort size and mean fact count. As filters are added, the cohort size decreases and the fact count increases. The time span filters have a greater impact than the data type filters. However, the RoutineVisit filter is unique in that it causes a decrease in the fact count. A possible explanation is that the RoutineVisit filter tends to select "healthy" patients, who have fewer data facts than sicker patients.

Filters applied to administrative claims data

The absence of certain data types or multiple visits in the EHR data can be the result of either incomplete data or good health. It is not possible to distinguish the 2 using EHR data alone. In contrast, when applying the filters to the claims data, we see the effects of the filters on a starting group of individuals whose data are complete. Note that the fraction of members in the claims data who pass the filters is the true positive rate (TPR) of the filters. (Although, because EHR and claims data are not entirely equivalent, these TPRs might not translate the same way to EHR data.) Table 2 shows that all members passed the AgeSex filters. However, similar to the EHR data, each of the remaining filters reduced the size of the cohort and increased the fact count, the fraction of female members (with the exception of NoSmallFactCount), and the fraction of members with diagnoses and routine visits. Without any filters, members in the claims data had a mean fact count of 291.7, were 51.6% female, and had a mean age group of 43.8 years. Using this as a reference point, we can estimate the magnitude of the biases caused by different filters. For example, of the 7.1 million members, the 2.9 million with laboratory tests had a mean 404.1 facts (+38.5%), were 55.6% female (+4.0%), and had a mean age group of 47.0 (+3.2 years).

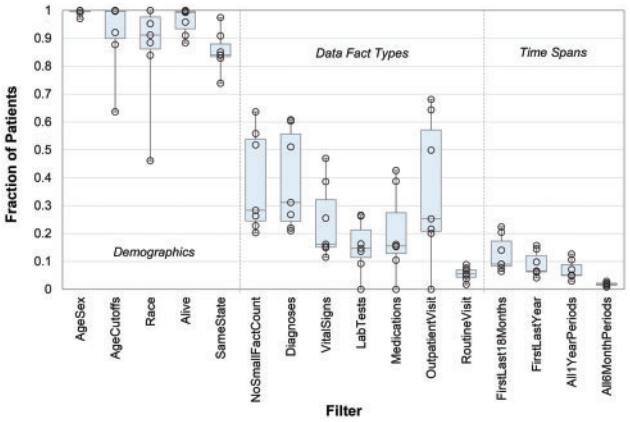

Figure 2. Fraction of patients at each of the 7 electronic health record sites who passed each of the 16 filters. Boxes indicate the median and quartiles.

Using the filters in practice

The optimal subset of filters depends on the particular research question and the extent to which potential biases are tolerable. For example, a study that does not require laboratory test results can ignore the LabTests filter to increase the cohort size and reduce a source of bias.

In the extreme case, where having fully complete data is essential, all 16 filters can be applied in order to minimize the chance of missing data. However, in the claims data, all 16 filters combined reduces the cohort size to only 722 885 members (10.2%), increases the fraction of female patients to 67.6% (+16.0%), and raises the mean fact count to 523.1 (+79.3%). Of the 12 million patients with EHR data, only 62 475 (0.5%) passed all filters. These patients were 70.0% female (+16.5%) and had a mean fact count of 678.4 (+973%).

The leadership at the Patient-Centered Outcomes Research Institute presented us with the challenge of identifying across the SCILHS sites a 1 million patient cohort with nearly complete data over a 4-year period. We wanted to minimize the biases in this SCILHS cohort and ensure that it had enough patients, although demographics and diagnoses are important for most use cases for the network. Therefore, we chose 8 of the 16 filters: all 5 demographic filters, NoSmallFactCount, Diagnoses, and

Table 2. The effects of filters applied to EHR and payor claims databases

Filter	Number of Patients	Fraction of Patients	Facts Per Patient	Fraction Female	Mean Age Group	Fraction of patients with...				
						Diagnoses	Routine Visits	Outpatient Visits	Inpatient Visits	Emergency Visits
Electronic Health Records (48 months at 7 sites)										
AllPatients	12 419 345	1.000	63.2	0.535	42.8	0.350	0.051	0.316	0.057	0.060
AgeSex	12 324 600	0.992	63.7	0.536	43.0	0.353	0.051	0.317	0.057	0.061
AgeCutoffs	11 366 647	0.915	68.6	0.537	43.1	0.376	0.055	0.335	0.061	0.066
Race	10 535 006	0.848	73.4	0.541	43.4	0.399	0.060	0.357	0.065	0.071
Alive	11 762 698	0.947	59.9	0.538	41.0	0.363	0.054	0.327	0.056	0.062
SameState	10 526 121	0.848	65.1	0.537	42.2	0.368	0.059	0.329	0.060	0.064
NoSmallFactCount	4 319 560	0.348	181.5	0.550	38.0	0.946	0.146	0.819	0.162	0.170
Diagnoses	4 352 033	0.350	179.0	0.550	37.9	1.000	0.146	0.826	0.162	0.170
VitalSigns	2 520 142	0.203	265.5	0.554	35.9	0.972	0.209	0.876	0.180	0.211
LabTests	2 032 271	0.164	325.5	0.569	43.1	0.976	0.235	0.836	0.267	0.177
Medications	2 676 609	0.216	252.0	0.560	42.3	0.909	0.172	0.814	0.217	0.142
OutpatientVisit	3 920 405	0.316	185.7	0.557	38.5	0.917	0.141	1.000	0.162	0.145
RoutineVisit	634 600	0.051	239.4	0.651	35.6	1.000	1.000	0.873	0.124	0.192
FirstLast18Months	1 515 146	0.122	328.5	0.580	41.9	0.978	0.272	0.901	0.191	0.211
FirstLastYear	1 082 160	0.087	379.4	0.586	43.4	0.981	0.309	0.905	0.199	0.219
All1YearPeriods	865 764	0.070	445.8	0.594	45.5	0.985	0.341	0.912	0.217	0.217
All6MonthPeriods	209 836	0.017	787.8	0.591	49.0	0.985	0.424	0.870	0.273	0.308
Payor Claims Data (members continually enrolled for 41 months)										
AllPatients	7 099 393	1.000	291.7	0.516	43.8	0.884	0.578			
AgeSex	7 099 339	1.000	291.7	0.516	43.8	0.884	0.578			
NoSmallFactCount	6 811 200	0.959	304.1	0.515	43.6	0.921	0.602			
Diagnoses	6 275 004	0.884	321.5	0.518	42.8	1.000	0.654			
LabTests	2 905 096	0.409	404.1	0.556	47.0	0.992	0.739			
Medications	2 995 556	0.422	339.9	0.534	45.0	0.891	0.614			
RoutineVisit	4 102 723	0.578	324.7	0.598	39.0	1.000	1.000			
FirstLast18Months	6 018 143	0.848	338.4	0.535	44.2	0.952	0.652			
FirstLastYear	5 825 623	0.821	345.7	0.539	44.4	0.952	0.657			
All1YearPeriods	5 596 228	0.788	356.5	0.545	44.7	0.951	0.665			
All6MonthPeriods	3 512 271	0.495	469.4	0.579	49.6	0.931	0.661			

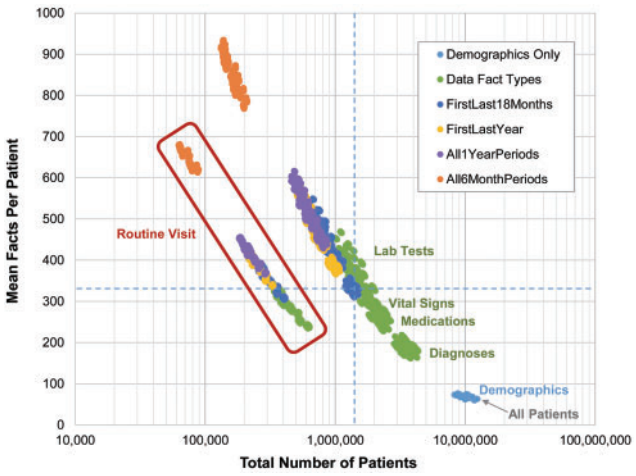

Figure 3. All $2^{16} = 65\,536$ filter combinations applied to the electronic health record data. The dotted blue lines indicate the number of patients and mean facts per patient for the filter combination selected for the SCILHS cohort. Light blue points are combinations of only demographic filters. Green points are filter combinations that include data fact type filters, but no time span filters. (Combinations that include the LabTests filter exclude the most patients, followed by VitalSigns, Medications, and Diagnoses.) The dark blue, yellow, purple, and orange points are combinations that include time span filters.

FirstLast18Months. These filters selected 1 335 571 patients, with a mean fact count of 319.7, 58.3% female, and a mean age group of 41.8 years (dotted blue lines in Figure 3). Compared to the claims data, the fact count was only 9.6% higher, there were 13.0% more female patients, and the mean age group was 2.0 years younger. Of the SCILHS cohort patients, 29.6%, 89.3%, 19.0%, and 22.2% had routine, outpatient, inpatient, and emergency room visits, respectively.

Only 5 of the 8 filters used for the SCILHS cohort are applicable to the claims dataset, since the data do not include members' race, whether they are alive, or the hospitals they visited. Those 5 filters selected 5 728 559 of the members in the claims data (TPR = 0.807). In other words, although the SCILHS cohort represents only a small fraction of the patients in the EHR data, the claims data suggest that most patients with complete data passed the filters. The members selected by the filters in the claims data had a mean fact count of 346.2, representing a +18.7% bias over the claims data as a whole. In contrast, the members in the claims data who did not pass the filters had very little data, with a mean fact count of only 64.0.

Comparing the SCILHS cohort and the claims data by age group

Figure 4 compares the SCILHS cohorts within each of the 7 EHR datasets, the full combined SCILHS cohort, and the claims data. The

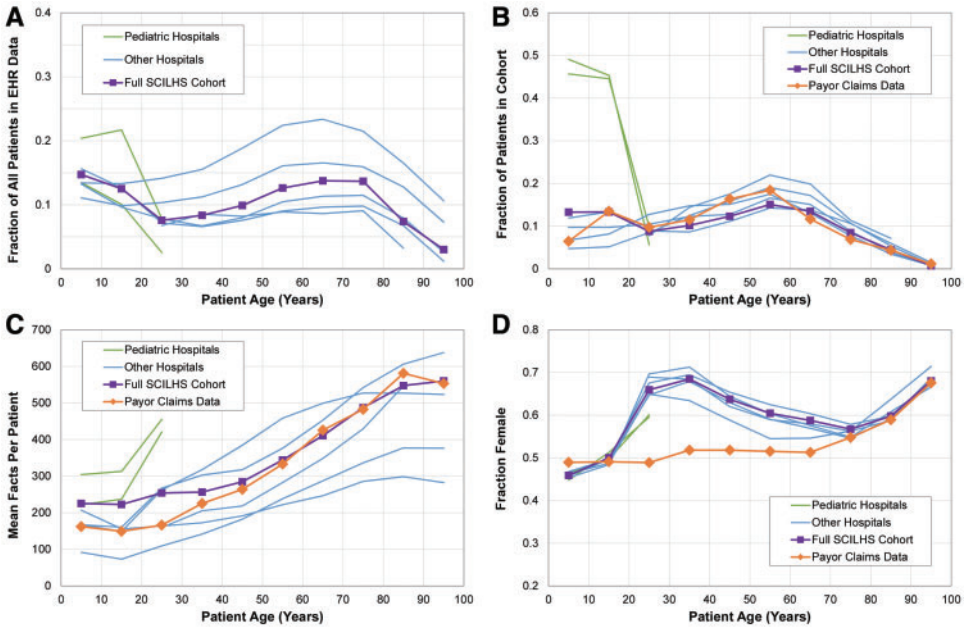

Figure 4. Age breakdowns for the SCILHS cohort. The top left graph shows (**A**) the fraction of all patients who were in the SCILHS cohort. The remaining graphs show (**B**) the age, (**C**) fact count, and (**D**) sex distributions of the SCILHS cohorts and the claims data. Each graph shows the 2 pediatric hospitals (green), the 5 mostly adult health care systems (blue), the 7 combined health care systems (purple squares), and the claims data (orange diamonds).

top left graph shows that in most cases, out of the entire EHR patient population, a larger fraction of patients in the 5-to-15-year-old and 45-to-65-year-old age groups were in the SCILHS cohorts than patients in the 25-to-35-year-old and 75-to-95-year-old age groups. The top right graph in Figure 4 shows that within the SCILHS cohorts, the 55-year-old age group was most common. The exceptions, of course, are the 2 pediatric hospitals, which only have patients <30 years old as a result of the AgeCutoff filer. The age distributions of the full SCILHS cohort and the claims data are similar, though there are about twice as many SCILHS patients in the 5-year-old age group.

The bottom 2 graphs in Figure 4 show 2 other differences between the SCILHS cohorts and claims data. The pediatric hospitals increase the overall fact count for patients <30 years old in the SCILHS cohort compared to the claims data, and the fraction of female patients in the SCHILS cohort exceeds that in the claims data for patients in the 25-to-65-year-old age group. In particular, nearly 70% of patients in the 35-year-old age group in the SCILHS cohort were female, compared to about an even number of female and male members in the claims data.

DISCUSSION

Ideally, investigators using EHR data could link patient records to claims data to verify data completeness. However, linked EHR and claims datasets are a relatively rare commodity, and our SCILHS use case is just 1 example where this was not a feasible option. Therefore, in this study we evaluated a more practical approach by looking at 16 simple heuristic filters, which can be combined in 65 536 different ways to generate cohorts of patients with varying degrees of data completeness and biases. For any given use case, some of the subsets might be more appropriate than others. For example, a clinical study seeking to recruit patients with hepatitis C needs a single visit where this diagnosis is recorded, but complete data over a time span may not be necessary. A study of hemoglobin A1c management in patients with diabetes who take insulin would need laboratory test results and medications, but not necessarily a million patients. In generating our SCILHS cohort, we presented a use case where we balanced the need for a large number of patients with complete data against a desire to minimize bias.

In the absence of a linked dataset of sufficient size, we used claims data in a couple ways to help indirectly measure biases caused by the filters. First, we showed that the filters have similar effects, in both EHR and claims data, on various characteristics of the patient population, such as age, gender, and fact count. Because we do not know for certain which patients have complete EHR data, we cannot determine the absolute bias introduced by the filters when applied to EHR data. However, we can calculate the bias for the claims data and assume it is similar for EHR data. Second, we were able to select a combination of filters for the SCILHS cohort that produced patient characteristics far more similar to the claims data than the overall EHR population. Thus, although this does not necessarily mean the data for the SCILHS cohort are complete, it does suggest that the filters are selecting patients whose data are closer to being complete.

It is important to remember that the EHR and claims datasets were mostly different patient populations, and that there are fundamental differences between EHR and claims data (eg, a layer of quality control may exist in the claims data that is not present in raw EHR data). Thus, we can only make rough qualitative comparisons between the 2 experiments, and in actuality the types of biases caused by the filters on EHR and claims data are unlikely to be exactly the same.

Despite the limitations, our findings were similar to previous research on EHR data completeness. For example, in a population of 3.9 million patients, Weiskopf reported that 97.8% and 99.6% of patients had date of birth and sex recorded, while only 44.5%,

29.3%, and 12.6% had, respectively, diagnoses, laboratory test results, and medications reported.[19] These are well within the ranges of the SCILHS sites. Only 0.6% of patients met all 4 of Weiskopf's data completeness definitions, which is similar to the 0.5% of patients who passed all 16 of our filters. The fact that so few patients met all criteria emphasizes the importance of taking data completeness into account when conducting EHR-based research, but it also highlights the need to be strategic in applying only the subset of filters necessary for a given study.

There are several future directions for this research:

1. *Direct validation using linked EHR and claims data*: Although this might not be possible with large populations, it could be easier to link data for a small subset of patients, in which case the true biases and ability of the filters to identify patients with complete data could be tested, but it might be difficult to generalize those results to the entire population.
2. *Alternative approaches to evaluating data completeness*: More sophisticated ways of combining filters are possible, such as probabilistic models that weigh some filters more than others and select patients who pass a probability threshold. Also, we only looked at broad categories of data, such as the presence or absence of diagnoses as a whole. Rusanov showed that fact count can vary greatly when specific types of diagnoses (eg, neoplasms) are recorded.[18] Thus, certain filters could have different types of biases, depending on the known characteristics of the patients.
3. *Guidelines for selecting the best data completeness goals:* Another area of future research is identifying categories of research questions where certain types of missing data are acceptable. For example, a patient might see a primary care physician at one hospital but receive care from a specialist at another hospital. Although neither hospital has a complete medical history of the patient, her EHR data at the second hospital would still be useful for a study focused on the subspecialty of interest. A related question is: When having missing data is unavoidable, how does that bias the results of a study based on those data?
4. *Use cases beyond research*: In addition to research, there are other use cases for identifying patients with complete data, for example helping at-risk organizations, such as accountable care organizations, define target markets.[21,22]

CONCLUSION

In conclusion, we evaluated 16 heuristic filters, individually and in combination, that check for data completeness in EHR data. While this is far from an ideal approach to solving the problem of missing data, the filters are examples of simple practical techniques that many investigators use to ensure that a patient cohort selected from EHR data has the necessary data elements required for a study. However, what we contribute here is a look at the extent to which these filters reduce the number of patients in the cohort and the biases the filters introduce. It is essential for investigators to be aware of these unintended effects of the filters as they design their studies and choose patient selection criteria.

FUNDING

This project was supported by Patient-Centered Outcomes Research Institute (PCORI) award CDRN130604608 for development of the National Patient-Centered Clinical Research Network, known as PCORNet. It was also supported by grants 5R01GM104303 from the National Institute of General Medical Sciences/National Institutes of Health (NIH), U54HG007963 from the National Human Genome Research Institute/NIH, and U01CA198934 from the National Cancer Institute/NIH. Disclaimer: The statements presented in this publication are solely the responsibility of the authors and do not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee, or other participants in PCORnet.

CONTRIBUTORS

GW, SM, and KM were responsible for the study concept and design; GW, WA, EB, JB, KF, KM, AT, and XZ provided access to data; GW performed the data analysis; and all authors contributed to interpretation of results and writing the manuscript.

COMPETING INTERESTS

The authors have no competing interests to disclose.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

For their assistance with local implementations of the filters, we would like to thank Galina Lozinski from Boston Medical Center, Parth Divekar from Cincinnati Children's Hospital Medical Center, Jeffrey Klann from Massachusetts General Hospital, Ronal Campbell from The University of Texas School of Biomedical Informatics at Houston, and Yaorong Ge, Kun Wei, and Jian Zhang from Wake Forest Baptist Medical Center.

REFERENCES

1. Devoe JE, Gold R, McIntire P, *et al*. Electronic health records vs Medicaid claims: completeness of diabetes preventive care data in community health centers. *Ann Fam Med*. 2011;9(4):351–58.
2. Hersh WR, Weiner MG, Embi PJ, *et al*. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51(8 Suppl 3):S30–37.
3. Heintzman J, Bailey SR, Hoopes MJ, *et al*. Agreement of Medicaid claims and electronic health records for assessing preventive care quality among adults. *J Am Med Inform Assoc*. 2014;21(4):720–24.
4. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. *Arch Int Med*. 2010;170:1989–95.
5. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. *AMIA Jt Summits Transl Sci Proc*. 2010;2010:1–5.
6. Finnell JT, Overhage JM, Grannis S. All health care is not local: an evaluation of the distribution of Emergency Department care delivered in Indiana. *AMIA Annu Symp Proc*. 2011;2011:409–16.
7. Lau EC, Mowat FS, Kelsh MA, *et al*. Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. *Clin Epidemiol*. 2011;3:259–72.
8. Wei WQ, Leibson CL, Ransom JE, *et al*. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc*. 2012;19(2):219–24.
9. Wei WQ, Leibson CL, Ransom JE, *et al*. The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *Int J Med Inform*. 2013;82(4):239–47.

10. Adler-Milstein J, Bates DW, Jha AK. Operational health information exchanges show substantial growth, but long-term funding remains a concern. *Health Aff (Millwood)*. 2013;32(8):1486–92.
11. Adler-Milstein J, Jha AK. Health information exchange among US hospitals: Who's in, who's out, and why? *Healthcare*. 2014;2(1):26–32.
12. Thorn SA, Carter MA, Bailey JE. Emergency physicians' perspectives on their use of health information exchange. *Ann Emerg Med*. 2014;63(3):329–37.
13. Yeager VA, Walker D, Cole E, *et al*. Factors related to health information exchange participation and use systems-level quality improvement. *J Med Syst*. 2014;38(8):78.
14. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *JAMA*. 2014;311(24):2479–80.
15. Mandl KD, Kohane IS, McFadden D, *et al*. Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): architecture. *J Am Med Inform Assoc*. 2014;21:615–20.
16. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. *J Am Med Inform Assoc*. 2014;21:576–77.
17. Selby JV, Lipstein SH. PCORI at 3 years: progress, lessons, and plans. *New Engl J Med*. 2014;370:592–95.
18. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak*. 2014;14:51
19. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46(5):830–36.
20. Murphy SN, Weber G, Mendis M, *et al*. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010;17(2):124–30
21. Lewis VA, Colla CH, Carluzzo KL, *et al*. Accountable Care Organizations in the United States: market and demographic factors associated with formation. *Health Services Res*. 2013;48:1840–58.
22. Scheffler RM, Shortell SM, Wilensky GR. Accountable care organizations and antitrust: restructuring the health care market. *JAMA*. 2012;307:1493–94.