Variation in Cohorts Derived from EHR Data in Four Care Delivery Settings

Susan Rea, PhD[1], Kent R. Bailey, PhD[2],
Jyotishman Pathak, PhD[2], Peter J. Haug, MD[1]
[1]Intermountain Healthcare, Salt Lake City, UT; [2]Mayo Clinic, Rochester, MN

Introduction

EHR data are desirable for secondary use in health research but are known to have inconsistencies. The context of their origination, or provenance, may affect the comparability of EHR data for secondary usage. We investigated suspected differences in demographic and comorbidity data and availability of information among four health care settings: ambulatory office visits, hospital inpatients, emergency visits, and visits for tests and other diagnostic and treatment procedures only. Descriptive comparative results suggest that cases accrued to a diabetes cohort in these settings differed on demographics, morbidity profiles and completeness of EHR data. The distribution of cases among the four settings also differed by provider geographic regions. These differences may reflect real differences in the health care services and documentation practices as well as differential patient access and utilization among the settings. This study demonstrates generalizable methods of classifying encounters in order to profile, compare and inform the use of secondary data across organizations.

Background

Hripcsak and Albers discuss the need for a better understanding of EHR data in context of the primary use environment where data were generated. This knowledge will enable innovative solutions to deal with known biases in secondary data.[1] Richesson, et al., acknowledged that phenotyping algorithms developed in particular health care settings will have unique biases in patient characteristics, utilization of services, and documentation practices that affect the performance of the algorithm in other settings.[2] Jensen, et al., describe the 'overfitting' and systematic bias in EHR data as a significant weakness in the pursuit of machine learning and prediction from secondary health care data. [3]

However, studies exposing specific effects of EHR data provenance were not found. Understanding these effects and their generalizability across organizations enhances the usability of secondary EHR data. This analysis was part of an ongoing collaborative study of the effects of heterogeneity of EHR data on the generalizability of a Type 2 diabetes mellitus (T2DM) phenotyping algorithm.[4] [5] This algorithm was selected as a use case to demonstrate the SHARPn data normalization pipeline.[6] The pipeline persists normalized secondary EHR data for analytic processing, using standard terminologies and detailed clinical element models (CEMs).[7] CEMs support rich provenance, or context, for each data element. This analysis shows differences in the profiles of patient cohorts related to the context of the setting of care. We used standardized administrative data to classify the settings of care so that the analysis can be replicated in other health care delivery organizations.

Methods

Intermountain Healthcare is a nonprofit, integrated health care system with 22 hospitals, 185 ambulatory clinics, and a full spectrum of health services such as home care, rehabilitation, laboratories and advanced trauma centers. Its services cover the state of Utah. Electronic health record (EHR) systems have been used since before 1983.[8] There are two administrative systems: records of patient encounters for professional services by the medical group, clinical laboratories, and other ambulatory services; and records of institutional encounters, including hospitals, hospital-based clinics, and other sites of care such as nursing home and home health. Longitudinal EHR and administrative data for secondary use are managed in the Enterprise Data Warehouse (EDW). These resources enable the study of provenance factors on a large heterogeneous repository of patient data. In this study, we focused on demographic and comorbidity features of a patient cohort with evidence of DM, compared by the type of practice settings they visited. Intermountain IRB approval was granted for this study.

Approximately 10% of 110,000 adult patients with evidence of DM during 2007 – 2011 were randomly sampled from the EDW (n = 10,426). One ICD-9-CM code for DM (250.*) in the encounter diagnosis records in either the institutional or professional services administrative systems was used as the selection criterion for a potential DM case, irrespective of type. Hospital discharge diagnoses and ambulatory encounter diagnoses in all ICD-9-CM

sequence positions were included. All study data were drawn from this same 5 year period. We used standard administrative coding of CMS *place of service* (POS) [9] for professional services, institutional patient types (inpatient, outpatient) and emergent arrival status, and Berenson-Eggers Type of Service (BETOS) [10] codes to classify all encounters into four health care delivery setting groups. The encounter settings used in the study were (1) face to face provider visits where evaluation and a medical diagnosis are expected to occur, (2) hospital stays, (3) hospital emergency room visits, and (4) encounters for tests and procedures only. The differences in characteristics of the patients were compared for those who had at least one provider evaluation visit; those who had no known evaluation visit but had either inpatient or, separately, emergency encounters; and those who had none of the previous types of encounters but had visits for tests and procedures. The four comparison groups are referred to as (1) *AMB*, (2) *IP not AMB*, (3) *ED not AMB*, and (4) *TP Only*.

Nationally standard administrative data were used to classify the setting groups in order to generate generalizable results. Only four POS codes were used to define eligible professional service encounters: 11 (office), 20 (urgent care facility), 22 (outpatient hospital) and 81 (independent laboratory). Institutional administrative systems must categorize encounters as *inpatient* or *outpatient* and whether *emergent*. The BETOS codes summarize Healthcare Common Procedure Coding System (HCPCS) codes into six major groupings: physician evaluation and management (E&M), physician procedures, imaging, laboratory tests, durable medical equipment, and other. All HCPCS, which include CPT4, codes available in either administrative system were mapped to BETOS codes. We used the combination of POS 11, 20, or 22 or institutional outpatient (non-emergent) and a BETOS code of M1A, M1B or M6 [11] (ambulatory E&M or consultation services) to define the 'AMB' group. 'IP not AMB' and 'ED not AMB' groups were based on institutional patient types only. The 'TP Only' group consisted of POS 81 or institutional outpatient (non-emergent) type and a BETOS code of P*, I* or T* (procedures, imaging, tests).

Diagnosis data were summarized in Clinical Classifications Software[12] single-level categories in order to reduce many ICD-9-CM codes into meaningful groupings for analysis. Several CCS categories that are known comorbidities to DM were selected to compare for this analysis: DM complications, hypertension, coronary heart disease and chronic renal failure. Patient deaths noted in the EDW through July, 2013, were used. Death data were updated from the Utah state records in April, 2013. Intermountain has assigned its facilities to physical regions of the state of Utah and southeastern Idaho. For this analysis, the regions were summarized further into the *urban* regions, the *rural* areas, and *mixed* regions – those having local access to health care facilities and resources but distant from the urban centers. Visit counts were summarized both to compare utilization across settings and as a proxy measure of the depth of information that may be expected in the EHR.

Patient demographics, visit counts, and comorbidities were described for the 4 groups to assess whether patients identified as diabetics retrospectively from EHR data appear to differ by the setting of care.

Results and Discussion

Table 1 shows the distribution of the 10,426 cases to the health care delivery setting groups. The majority of patients with a DM diagnosis code in the 5 year period were seen at least once for medical evaluation (77%). The 'AMB' group showed a slightly higher proportion of women as well as much higher average visit counts. The 'IP not AMB' group was older, on average. The number of visits varies among the groups. There were 112 cases (1.1% of the study cohort) that had no visits classified into the groups used for comparison. The unclassified cases contained a higher proportion of men than other groups and even less visit history. The encounters for these cases were reviewed using additional administrative data. They comprised lab or imaging (40%), hospital general clinic visits (31%), hospital specialty clinic visits (17%), and professional services in other settings (13%). About half of these could be classified by incorporating local administrative data into the classification methods, but the intent was to use standardized administrative data to generate the groupings.

The distribution of setting groups in each region type is shown in Figure 2. We measured the percentage of all encounters in each region type that were classified to each setting. Each region's four setting percentages total 100. Regions may share cases so these data can only suggest a trend in access to care in these settings. The trend is for a higher proportion of cases in the 'AMB' setting in urban versus rural regions (75% v. 52%) and a higher proportion of DM cases in the 'ED not AMB' setting (23% v. 11%) and the 'TP Only' setting (18% v. 9%) in the rural versus urban regions.

VISIT SETTING	# CASES	% OF COHORT*	AVG AGE	% FEMALE	AVG VISIT COUNT	AVG DM VISIT COUNT
AMB	7984	76.6	57.6	51.4	31.7	9.6
IP not AMB	972	9.3	61.9	47.3	8.4	3.0
ED not AMB	1192	11.4	56.9	47.7	7.6	2.9
TP only	583	5.6	57.8	46.8	5.2	2.0
unclassified	112	1.1	58.1	41.1	2.4	1.5
* IP and ED rows share 417 cases.						
** Known deaths updated July, 2013.						

Table 1. Distribution of cases and characteristics by setting

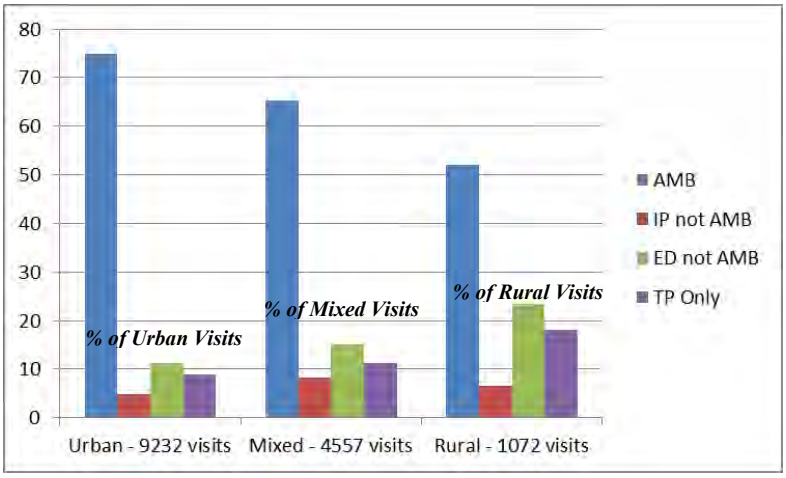

Figure 1. Percent of cases accrued to setting groups by regions

The volume and settings of visits were compared across the setting groups (Table 2). The 'IP not AMB' group had more inpatient encounters than other groups. We cannot compare ED visits between the 'IP not AMB' and the 'ED not AMB' groups because standard hospital coding assigns inpatient status to patients admitted from the ED. Table 2 confirms the visit data in Table 1 showing few encounters over 5 years for all settings other than 'AMB'. A 43% rate of hospitalization in the 'AMB' group was similar to a rate of 42.6% previously reported for 18,404 diabetes cases.[13]

VISIT SETTING	% HAVING IP VISITS	AVG# IP VISITS PER CASE	% HAVING ED VISITS	AVG# ED VISITS PER CASE	% HAVING TP VISITS	AVG# TP VISITS PER CASE
AMB	43	1	50	2	94	14
IP not AMB	100	2	43	1	63	3
ED not AMB	35	1	100	2	55	3
TP only	0	0	0	0	100	4

Table 2. Visit volume across settings by setting groups

Figure 2 shows a potential problem with mixing the cases drawn from these settings in a secondary research cohort. The proportion of documented comorbidities and known death per case are shown for each setting group. The mortality rates are highest in the 'IP not AMB' group and suggest this group would have more morbidity. Although the proportion of cases with coronary heart disease and chronic renal failure are somewhat higher than the 'AMB' group (34% v. 27% and 21% v. 15%), the hypertension and DM complications proportions follow the decreasing trend for comorbidities for all settings compared to the 'AMB' group. These data suggest there may be less documentation for comorbidities related to the coding practices in the non-'AMB' settings or related to the setting groups' lower average visit counts shown in Table 1.

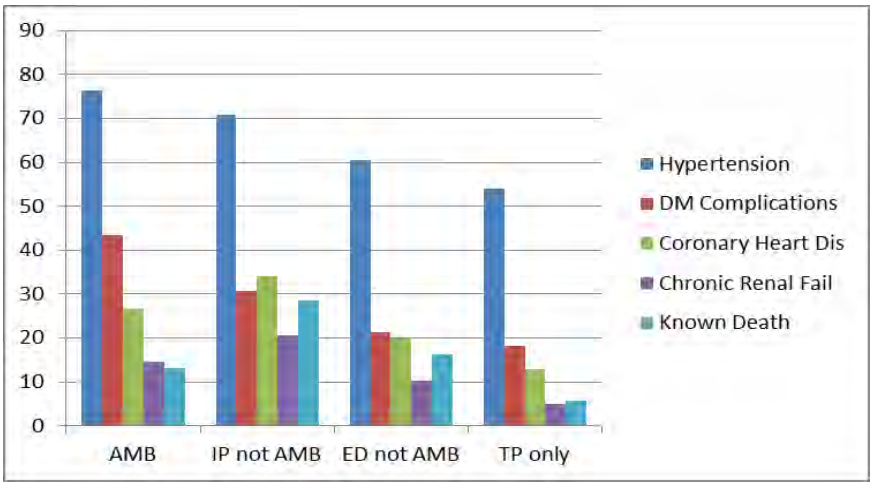

Figure 2. Proportion of cases having documented comorbidities or death

Conclusions

These data suggest demographic, morbidity, mortality and data availability differences in DM cases by provider setting where they might be identified by a DM phenotyping algorithm. The comorbidity profile for cases in setting groups having fewer encounters probably reflects more missing data rather than lower disease burden. Intermountain, like other health care delivery organizations, has a unique mix of provider settings and population access to the settings. This descriptive study was intended to discover and describe setting data differences that might affect the comparability and usability of secondary EHR data. We also demonstrated standard methods to classify cases into the setting groups so that similar profiling may be used to compare cohorts across organizations. Although administrative data, included in an EHR, were used for this study, they signal downstream effects in the clinical observations recorded in these settings. The provenance of clinical observations in the EHR also conveys important primary use contextual information that can inform the comparability of secondary data aggregated for research.

Limitations and Further Research

The classification of institutional outpatient data may be improved by the use of standard revenue codes in addition to or in place of BETOS codes. These data were a 'snapshot' of cases over a 5 year time period, with consequent loss of case data outside these bounds. The data were sampled from one organization and results do not reflect other provider organizations. The focus of our research is not the explanation of specific inconsistencies in EHR data, but rather to contribute generalizable methods to expose data quality issues and remediation opportunities. Further

research of data provenance and data heterogeneity across provider organizations and the effects on the accuracy of specific phenotyping algorithms is underway.

Acknowledgements
This manuscript was made possible by funding from the Strategic Health IT Advanced Research Projects (SHARP) Program (90TR002) administered by the Office of the National Coordinator for Health Information Technology. The contents of the manuscript are solely the responsibility of the authors.

References
1 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013 Jan 1;**20**(1):117-21.
2 Richesson RL, Rusincovitch SA, Wixted D, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc. 2013 Sep 11.
3 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature reviews Genetics. 2012 Jun;**13**(6):395-405.
4 Bailey KR, Rea S, Wood-Wentz CM. Extracting Data from EHRs for Algorithm Implementation at Two Comprehensive Care Institutions Data Sources, Pitfalls, Uncertainties. AMIA Joint Summit on Clinical Research Informatics; 2013; San Francisco, CA; 2013.
5 Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc. 2012 Mar-Apr;**19**(2):212-8.
6 Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. Journal of Biomedical Informatics. 2012;**45**(4):763-71.
7 Coyle JF, Mori AR, Huff SM. Standards for detailed clinical models as the basis for medical data exchange and decision support. Int J Med Inform. 2003 Mar;**69**(2-3):157-74.
8 Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP system. Journal of medical systems. 1983 Apr;**7**(2):87-102.
9 Place of Service Code Set. 2012 Nov, 2012 [cited July, 2013]; Available from: http://www.cms.gov/Medicare/Coding/place-of-service-codes/Place_of_Service_Code_Set.html
10 Berenson-Eggers Type of Service (BETOS). 2012 Dec 6, 2012 [cited June 1, 2013]; Available from: http://www.cms.gov/Medicare/Coding/HCPCSReleaseCodeSets/BETOS.html
11 Berenson-Eggers Type of Service (BETOS) Codes. 2012 [cited Jul, 2013]; Available from: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareFeeforSvcPartsAB/downloads/betosdesccodes.pdf
12 Clinical Classifications Software (CCS) for ICD-9-CM. 2013 Mar 28, 2013 [cited June 1, 2013]; Available from: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp
13 Robbins JM, Thatcher GE, Webb DA, Valdmanis VG. Nutritionist Visits, Diabetes Classes, and Hospitalization Rates and Charges: The Urban Diabetes Study. Diabetes care. 2008 April 1, 2008;**31**(4):655-60.