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

Wei-Qi Wei,[1,2] Cynthia L Leibson,[3] Jeanine E Ransom,[2] Abel N Kho,[4] Pedro J Caraballo,[5] High Seng Chai,[2] Barbara P Yawn,[6] Jennifer A Pacheco,[7] Christopher G Chute[2]

► Additional appendices are published online only. To view these files please visit the journal online (http://jamia.bmj.com/content/19/2.toc).

[1]Institute for Health Informatics, University of Minnesota, Twin Cities, Minnesota, USA
[2]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA
[3]Division of Epidemiology, Mayo Clinic, Rochester, Minnesota, USA
[4]Divisions of General Internal Medicine and Biomedical Informatics, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA
[5]Division of General Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA
[6]Department of Research, Olmsted Medical Center, Rochester, Minnesota, USA
[7]Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

Correspondence to
Dr Christopher G Chute, Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA; chute@mayo.edu

Received 13 September 2011
Accepted 21 December 2011
Published Online First 16 January 2012

ABSTRACT

Objective To evaluate data fragmentation across healthcare centers with regard to the accuracy of a high-throughput clinical phenotyping (HTCP) algorithm developed to differentiate (1) patients with type 2 diabetes mellitus (T2DM) and (2) patients with no diabetes.

Materials and methods This population-based study identified all Olmsted County, Minnesota residents in 2007. We used provider-linked electronic medical record data from the two healthcare centers that provide >95% of all care to County residents (ie, Olmsted Medical Center and Mayo Clinic in Rochester, Minnesota, USA). Subjects were limited to residents with one or more encounter January 1, 2006 through December 31, 2007 at both healthcare centers. DM-relevant data on diagnoses, laboratory results, and medication from both centers were obtained during this period. The algorithm was first executed using data from both centers (ie, the gold standard) and then from Mayo Clinic alone. Positive predictive values and false-negative rates were calculated, and the McNemar test was used to compare categorization when data from the Mayo Clinic alone were used with the gold standard. Age and sex were compared between true-positive and false-negative subjects with T2DM. Statistical significance was accepted as $p<0.05$.

Results With data from both medical centers, 765 subjects with T2DM (4256 non-DM subjects) were identified. When single-center data were used, 252 T2DM subjects (1573 non-DM subjects) were missed; an additional false-positive 27 T2DM subjects (215 non-DM subjects) were identified. The positive predictive values and false-negative rates were 95.0% (513/540) and 32.9% (252/765), respectively, for T2DM subjects and 92.6% (2683/2898) and 37.0% (1573/4256), respectively, for non-DM subjects. Age and sex distribution differed between true-positive (mean age 62.1; 45% female) and false-negative (mean age 65.0; 56.0% female) T2DM subjects.

Conclusion The findings show that application of an HTCP algorithm using data from a single medical center contributes to misclassification. These findings should be considered carefully by researchers when developing and executing HTCP algorithms.

BACKGROUND AND SIGNIFICANCE

Subject selection—the process of identifying patients with specific clinical characteristics—is an essential component of clinical studies. Accurate selection consumes considerable time and effort to gather, abstract, and review medical charts, and it is often the rate-limiting step in clinical research.[1] Recently, the increased adoption of electronic medical record (EMR) systems has provided researchers with an advanced tool to improve this inefficient process.[2] By leveraging the machine-processable content through an EMR system, clinical researchers can develop a high-throughput clinical phenotyping (HTCP) algorithm (a set of inclusion and exclusion criteria for identifying patients with specified characteristics), execute the algorithm against already existing data within an EMR system, and rapidly obtain a large pool of eligible study subjects.[3–6]

The Electronic Medical Records and Genomics (eMERGE) Network,[5] a national consortium funded by the National Human Genome Research Institute, has devoted substantial efforts to exploring the possibility of leveraging EMRs as resources for subject selection. The eMERGE I Network consisted of five national leading academic medical centers: Mayo Clinic, Rochester, Minnesota; Northwestern University Medical Center, Chicago, Illinois; Vanderbilt University Medical Center, Nashville, Tennessee; Marshfield Clinic in Wisconsin, Marshfield, Wisconsin; and the Group Health Cooperative with the University of Washington, Seattle, Washington. One of its primary goals was to develop HTCP algorithms for identifying subjects suitable for genotype- and phenotype-associated studies. In order to ensure that an algorithm is transportable and that various institutions can execute it to obtain reliable outputs, each algorithm developed in the eMERGE Network was proposed, reviewed, and validated by domain experts across participating medical centers.

The HTCP approach of leveraging EMR data for subject selection is appealing because it offers increased efficiency while reducing the large amount of manual detail work that is required. We hypothesize that results of HTCP are more accurate if all medical data for every patient are available for review. However, the ability to capture all of

a patient's medical data is limited when patients are seen by multiple healthcare centers. A recent study indicates that, of the nearly 3.7 million patients who sought treatment in acute care settings in Massachusetts during a 5-year period, over 30% visited more than one hospital and 1%—or 43 794 patients—visited five or more hospitals during the study period.[7] Similar findings on multiple healthcare centers for primary care visits were reported by Smith *et al*.[8] The resultant data fragmentation across healthcare centers leads to incomplete data from any one EMR when researchers execute an HTCP algorithm at a single medical center. The absent data could be crucial in qualifying or disqualifying a study subject and could cause subject selection errors.

Previous studies of the effect of data fragmentation on clinical outcomes suggested that data fragmentation wasted valuable medical resources and could adversely affect treatment outcomes.[8–13] Cox and his colleagues[14] investigated the influence of missing data and demonstrated that subjects with missing data differed significantly in terms of variables crucial to the study outcome and that distortion led to biased results.

To our knowledge, the impact of data fragmentation across healthcare centers on an HTCP algorithm has not been explicitly investigated. The present study evaluated the effect of data fragmentation on an HTCP algorithm developed within the eMERGE Network for specifying patients with type 2 diabetes mellitus (T2DM).

THE EMERGE T2DM ALGORITHM

T2DM is a multiple gene-related chronic disease that poses an enormous public health burden.[15] As provided in detail elsewhere (unpublished material, Wei W, 2011),[16] the eMERGE T2DM algorithm is EMR based and was developed by researchers from Northwestern University and enhanced by other participating institutes within the eMERGE Network. The primary goal of this algorithm is to maximize the positive predictive value (PPV) or the precision of identifying 'T2DM subjects', a term used herein to mean patients with T2DM, and to avoid confounding by inclusion as subjects individuals without any type of diabetes mellitus (DM) or individuals with type 1 DM (T1DM). With respect to unaffected subjects (herein termed 'non-DM subjects'), the goal of the algorithm is to maximize the PPV of identifying individuals with no DM, excluding even those at risk of DM which has not yet manifested itself (ie, pre-DM).

Previous evidence has suggested that ICD-9-CM (*International Classification of Diseases*, Ninth Revision, Clinical Modification) codes alone would not provide enough accuracy to identify patients with DM.[17 18] More importantly, T2DM subjects identified using only diagnosis codes could be contaminated with T1DM subjects because many patients are assigned the code for 'diabetes mellitus, unspecified type' and some patients with T2DM diagnosis codes are actually T1DM subjects who have been wrongly assigned a code for T2DM. To avoid such potential misclassification, the algorithm developers supplemented the use of diagnosis codes with relevant laboratory results and medication prescriptions (figures 1 and 2).

Previous evaluation studies indicated that the algorithm achieved 98% and 100% PPVs for identification of T2DM subjects and non-DM subjects, respectively, compared with clinician review.[16] However, both the EMR data and the records that were reviewed came from the same medical center. Thus the effect of data fragmentation across healthcare centers on its performance is still unknown. We chose to evaluate the effect of data fragmentation across healthcare centers on the basis of this algorithm because it involves virtually all structured EMR data (ie, diagnosis, laboratory values, and medication) and has demonstrated high accuracy within a single medical center.

MATERIALS AND METHODS

Study setting

This is a population-based medical records study. It was conducted in Olmsted County, Minnesota (2010 census =144 248).

Data resources

The study took advantage of Rochester Epidemiology Project (REP) resources.[19] REP is a medical records-linkage system for all residents of Olmsted County, which has been continuously funded by National Institutes of Health since 1966. Population-based studies using REP resources are afforded because Rochester, the county seat, is geographically isolated (approximately 136 km from the nearest urban center) and home to Mayo Clinic, one of the world's largest medical centers. Thus >95% of medical care received by County residents is provided by either Mayo Clinic, with its two affiliated hospitals, or Olmsted Medical Center (OMC), a second group practice, with its affiliated hospital.[20] The REP maintains a unique identifier for each Olmsted County resident over time and across

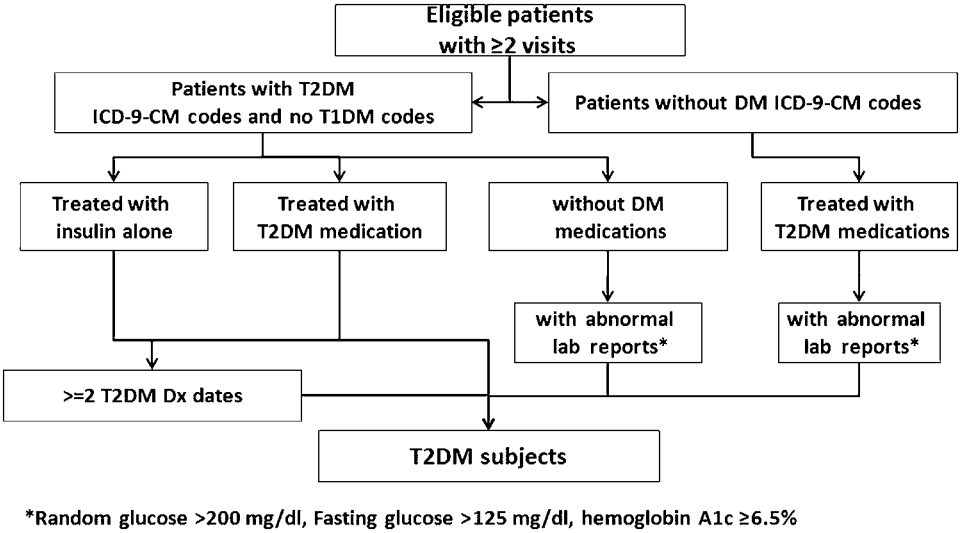

Figure 1 The eMERGE algorithm for identifying T2DM subjects. DM, diabetes mellitus; Dx, diagnosis; eMERGE, Electronic Medical Records and Genomics; ICD-9-CM, *International Classification of Diseases*, 9th Revision, Clinical Modification; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus.

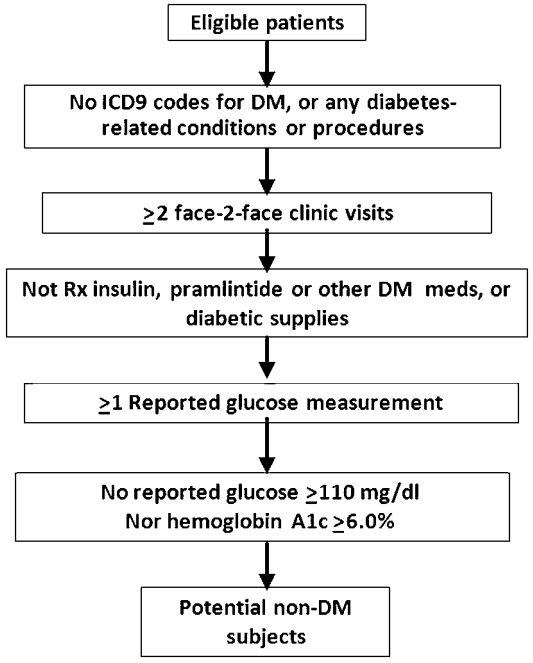

Figure 2 The eMERGE algorithm for identifying non-DM subjects. DM, diabetes mellitus; eMERGE, Electronic Medical Records and Genomics; ICD9, *International Classification of Diseases*, 9th Revision; Rx, prescription.

healthcare centers, and each resident's clinical data from virtually all sources of medical care (hospital inpatient, hospital outpatient, emergency department, office, and nursing home visits) can be combined for approved clinical research.[19 20]

Eligible subjects

The study was approved by the Mayo Clinic and the OMC Institutional Review Boards. We first used the REP census[19] to identify all unique individuals residing in Olmsted County in 2007. Persons who refused authorization for use of their medical record in research at either OMC or Mayo Clinic (typically, <5%[21]) were excluded. To be eligible for this study, subjects had to be Olmsted County residents and to have had at least one encounter at Mayo Clinic from January 1, 2006 through December 31, 2007 and also at least one encounter at OMC from January 1, 2006 through December 31, 2007.

EMR data

We obtained 2 years of EMR data for eligible patients (2006 and 2007) from OMC and Mayo Clinic separately. We searched administrative claims data to determine the presence (or absence) of DM-relevant ICD-9-CM codes (see online appendix (ICD_codes)). Outpatient laboratory data for DM-relevant tests were reviewed to determine whether a subject had an abnormal value. The sources of medication data were electronic outpatient prescription databases. One author (PJC), a licensed internal medicine physician with a focus on diabetes, manually reviewed the databases and provided a list of generic drug names, brand names, synonyms, and abbreviations for DM-relevant medications (see online appendix (drug list)). We searched the medication data for the terms on the list to determine whether or not a patient had been prescribed any such medications.

Data analysis

We first executed the eMERGE T2DM algorithm (figures 1 and 2) on EMR data combined from both OMC and Mayo Clinic EMR systems. The categorization of eligible patients as 'T2DM subjects' and 'non-DM subjects' using data from both EMR systems was considered the gold standard in this study. We then executed the algorithm using Mayo Clinic EMR data alone; for T2DM subjects and non-DM subjects separately, we calculated the number of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) against the gold standard. We also estimated PPVs and false-negative rates (FNRs) to evaluate misclassification errors caused by data fragmentation across healthcare centers.

The numerator for PPV is the number of TP subjects—that is, those identified using EMR data from both centers who were also identified as such using EMR data from Mayo Clinic alone. The denominator for PPV is the sum of TP subjects plus the number of FP subjects, with FP subjects defined as subjects categorized as not subjects by the gold standard but as subjects by EMR data from Mayo Clinic alone.

The numerator for FNR is the number of subjects categorized as subjects using the gold standard, but who were categorized as not subjects using EMR data from Mayo Clinic alone. The denominator for FNR is the number of FN subjects plus the number of TP subjects.

The McNemar test[22] was used to analyze whether the categorization that resulted from the use of EMR data from two centers differed from the categorization when Mayo Clinic EMR data alone were used. The distributions of two commonly used epidemiological characteristics (age and sex) were compared between TP T2DM subjects and FN T2DM subjects to estimate whether T2DM subjects falsely excluded because of data fragmentation were statistically different from identified T2DM subjects. Comparison of the mean age between two groups was performed with the t test. Comparison of sex proportions was performed with the χ^2 test. Statistical significance was accepted when p was <0.05. All data are presented as mean and SD. Statistical analysis was performed with R for Windows software V.2.11.1.[23]

RESULTS

Of 139 654 Olmsted County residents in 2007, 12 740 (9.1%) had at least one encounter at the Mayo Clinic from January 1, 2006 through December 31, 2007 and at least one encounter at OMC within the same time frame (table 1). These 12 740 residents were eligible for the present study.

T2DM subject identification

Of the 12 740 eligible subjects, 6.0% (765) met eMERGE T2DM algorithm inclusion criteria for T2DM subjects when their combined Mayo Clinic and OMC EMR data were used (table 2). These 765 subjects were considered true T2DM subjects for this study.

By comparison, 540 patients were identified as T2DM subjects when their Mayo Clinic EMR data alone were used; 513 were TP and 27 were FP (table 2). The PPV was 95% (513/540). The other 252 true T2DM subjects were FN (ie, incorrectly

Table 1 Demographic characteristics of Olmsted County residents and eligible patients

Characteristic	Value
Olmsted county residents in 2007 (n=139 654)	
Age (years), mean (SD)	35.8 (22.9)
Female sex, %	53.2
Eligible subjects (n=12 740)	
Age (years), mean (SD)	40.9 (23.0)
Female sex, %	54.9

Table 2 Categorization of eligible patients as T2DM and non-DM subjects

Data source	TPs, N	FPs, N	TNs, N	FNs, N	Sensitivity (TP/(TP+FN)),%	Specificity (TN/(FP+TN)), %	PPV (TP/(TP+FP)), %	FNR (FN/(TP+FN)), %	p Value*
T2DM subject									
Mayo Clinic alone	513	27	11 948	252	67.1	99.8	95.0	32.9	<0.001
Mayo Clinic + OMC	765	0	11 975	0	100	100	100	0	
Non-DM subject									
Mayo Clinic alone	2683	215	8269	1573	63.0	97.5	92.6	37.0	<0.001
Mayo Clinic + OMC	4256	0	8484	0	100	100	100	0	

*p Value for comparison of categorizations between Mayo Clinic alone and Mayo Clinic + OMC.
DM, diabetes mellitus; FNR, false-negative rate; FN, false negative; FP, false positive; OMC, Olmsted Medical Center; PPV, positive predictive value; TN, true negative; TP, true positive; T2DM, type 2 diabetes mellitus.

excluded when Mayo Clinic EMR data alone were used). The FNR was 32.9% (252/765). We found differences in the mean age (p=0.012) and sex proportion (p=0.004) between the group of 513 correctly identified T2DM subjects (62.1 (15.2) years; female to male ratio, 230:283) and the group of 252 missed T2DM subjects (65.0 (14.7) years; female to male ratio, 141:111). The McNemar test also indicated a difference between the categorization with EMR data from both centers and the categorization with data from Mayo Clinic alone (p<0.001).

With respect to which eMERGE inclusion/exclusion criteria (see figure 1) accounted for the misclassification of T2DM subjects with Mayo Clinic EMR data alone, all 27 FP T2DM subjects and 111 of the 252 (44%) FN T2DM subjects resulted from incomplete diagnosis codes at Mayo Clinic (table 3). Incomplete medication data at Mayo Clinic led to 75 (30%) FN T2DM subjects; an additional 53 (21%) FN T2DM subjects resulted from having only one encounter at Mayo Clinic 2006–2007 (the algorithm required at least two encounters). The remaining 13 (5%) FN T2DM subjects had abnormal laboratory results missing at Mayo Clinic.

Non-DM subject identification

With EMR data from both OMC and Mayo Clinic, 4256 subjects were identified by the algorithm as non-DM subjects (table 2). These were considered gold-standard non-DM subjects for this study.

With EMR data from Mayo Clinic alone, 2898 eligible subjects were categorized as non-DM subjects. However, only 2683 were TP, 215 were FP, and 1573 were FN (ie, incorrectly excluded as non-DM subjects when EMR data from a single medical center were used). The PPV and FNR were 92.6% (2683/2898) and 37.0% (1573/4256), respectively (table 2). Statistical analysis indicated a difference between the categorization with data from the two healthcare centers and that with data from Mayo Clinic alone (p <0.001).

With respect to which eMERGE inclusion/exclusion criteria (see figure 2) accounted for the misclassification of non-DM subjects with Mayo Clinic EMR data alone, incomplete laboratory data contributed to 135 (63%) FP non-DM subjects and 1074 (68%) FN non-DM subjects (table 3). Incomplete diagnosis codes contributed to another 73 (34%) FP non-DM subjects, and 499 (32%) of FN non-DM subjects resulted from having fewer than two encounters at Mayo Clinic 2006–2007.

Table 3 Number of, and reasons that contributed to, FP and FN results

Subjects	n (%)	Reasons that contributed to FP or FN results
FP		
T2DM subjects	27 (100)	Incomplete data for T1DM diagnosis
Non-DM subjects	73 (34)	Incomplete data for DM-relevant diagnosis
	7 (3)	Incomplete use history of antidiabetes medication or supplies
	135 (63)	Absence of laboratory results
FN		
T2DM subjects	53 (21)	<2 visits at Mayo Clinic between January 1, 2006 and December 31, 2007
	111 (44)	Incomplete data for T2DM diagnosis
	75 (30)	Incomplete treatment history of antidiabetes medication or supplies
	13 (5)	Absence of laboratory results
Non-DM subjects	499 (32)	<2 visits at Mayo Clinic between January 1, 2006 and December 31, 2007
	1074 (68)	Absence of laboratory results

DM, diabetes mellitus; FN, false negative; FP, false positive; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus.

DISCUSSION

Current clinical research is limited by a labor-intensive subject selection process, which has become a formidable obstacle to conducting broad and deep studies and drawing powerful conclusions. An HTCP algorithm leverages machine-processable EMR data, improving such inefficiency. Oftentimes, a patient is seen by multiple medical centers, and thus a single medical center does not have the patient's complete medical data when executing an algorithm. To our knowledge, how this data fragmentation across healthcare providers affects the accuracy of an HTCP algorithm has not been previously investigated. Such an investigation is difficult to conduct because it requires accessing multiple EMRs from heterogeneous sources at multiple medical centers. By taking advantage of the REP, we accomplished such a novel demonstration.

When using the combined Mayo Clinic and OMC EMR data for the 12 740 eligible subjects, 6.0% (765) met eMERGE T2DM algorithm inclusion criteria for T2DM subjects (table 2). This percentage is slightly lower than the prevalence of DM for all age groups in the USA (8.3%)[24] because not all Olmsted County residents were tested for DM in the 2 years of the study.

Our results, combined with findings from other studies,[8 14] show the advantage of access to more complete data for clinical research. In the present study, data fragmentation across healthcare centers resulted in incomplete data for any one EMR when the eMERGE T2DM algorithm was executed in Olmsted County, and that incompleteness substantially decreased the algorithm's accuracy.

For T2DM subject identification, we found categorization differences with data from both centers relative to the use of data from any one alone. The differences were mainly the result of a large proportion of FN T2DM subjects (n=252; FNR, 32.9%). The 252 FN T2DM subjects differed with respect to age and sex distribution from the 513 TP T2DM subjects. This difference suggests that, for age/sex-matched designs, matching could be skewed when HTCP algorithms are applied to EMR data from a single medical center. Even though the eMERGE T2DM

algorithm is reported to achieve 98% for identification of T2DM subjects compared with clinician review,[16] we still identified 27 (5.0%) FP T2DM subjects because of data fragmentation across healthcare centers.

For non-DM subject identification, we also found categorization differences using data from both centers relative to using data from any one alone. The differences were mainly the result of a large proportion of FN non-DM subjects (n=1573; FNR, 37.0%). Even though the eMERGE T2DM algorithm is reported to achieve 100% PPVs for identification of non-DM subjects compared with clinician review,[16] we still identified 215 (7.4%) FP non-DM subjects because of data fragmentation across healthcare centers.

An incomplete diagnosis is the main reason for FP errors and accounted for all FP T2DM subjects. Absent laboratory results and incomplete diagnosis led to the majority of FP non-DM subjects. FNs were caused by the incompleteness of diagnosis, laboratory values, or prior medications. We also found that 53 subjects (21%) and 499 subjects (32%) were missed because they had made fewer than two clinical visits during the study period. As the time frame we used was 2 years, which is broader than the recommended frequency of T2DM visits (3–6 months),[24 25] these insufficient clinical visits must have resulted from data fragmentation across centers as well.

The misclassification errors caused by data fragmentation could lead to sampling bias and risk serious distortions in the findings of resulting studies.[26] These outcomes should be carefully considered by clinical researchers when developing or executing an algorithm. The ultimate solution for the data fragmentation problem is integrating EMR systems across various healthcare centers. However, to achieve such an ambitious goal, not only do serious technological challenges exist, but also complex ethical issues need to be addressed. Some ONC (the Office of the National Coordinator) funded Beacon projects prototype this issue.[27]

Clinical narratives (unstructured clinical data) document a patient's detailed description about diseases that may contain data from other healthcare centers. This additional information can be extracted by using natural language processing techniques and turned into normalized data for further analysis using other advanced techniques—for example, data mining.[28] Then, discovered patterns could be reviewed and adopted in subject selection criteria. This approach may work with the caveat that additional data must be relevant for the condition under study. Our previous work, along with other studies, has shown its potential for subject selection tasks.[6 29–32]

Several issues about this study design should be considered when interpreting the findings. Because of unavoidable random or systematic errors (eg, physician experience, communication quality between the patient and the clinician, and coding quality), it is extremely difficult to obtain a patient's actual condition or the true gold standard.[33] The manual effort required to validate the distinction between T1DM and T2DM obtained using the algorithm against medical review requires information at the time of DM onset[34] and was beyond the scope of the present study. In this study, our gold standard was based on classifications using 2 years of EMR data from two major healthcare centers in Olmsted County. Because most Olmsted County residents receive their healthcare at these two healthcare centers and the observation window we chose is much broader than the recommended frequency of T2DM visits, this is a pragmatic gold standard for this study.

Our results may not generalize to large metropolitan areas. Our study setting is a sparsely populated, relatively isolated county in southeastern Minnesota. The residents of Olmsted County have fewer options for healthcare centers than people living in a large metropolitan area. Thus the misclassification errors that we found by comparing the selected categorizations are most likely smaller than in a usual situation. Also, this study focuses on how HCTP is affected by incomplete data due to data fragmentation across healthcare centers alone. It does not investigate the impact of incomplete data due to other factors, for example, insufficient longitudinal data, which is a topic for another study (unpublished material, Wei W, 2011). In addition, the algorithm scope of our study is limited to the eMERGE T2DM algorithm alone. For a more complete evaluation of the impact of data fragmentation on an HTCP algorithm, this study needs to be repeated at different geographic locations under various periods of observation on a wide spectrum of HTCP algorithms.

CONCLUSION

This study, to our knowledge, is the first attempt to assess the impact of data fragmentation on an HTCP algorithm across multi-institution EMRs. Our results show that data fragmentation across healthcare centers causes misclassification errors of an HTCP algorithm. This risk should be carefully considered by clinical researchers when developing or executing an HTCP algorithm.

Funding This study was supported by the Biomedical Informatics and Computational Biology Traineeship Program, the University of Minnesota, and the eMERGE project, NIH U01 HG04599.

Competing interests None.

Patient consent Obtained.

Ethics approval Mayo Clinic and OMC IRBs.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. *Delay No More: Improve Patient Recruitment and Reduce Time to Market in the Pharmaceutical Industry*. https://www-935.ibm.com/services/sg/index.wss/ibvstudy/igs/x1014229?cntxt=x1013529
2. **Wilke RA,** Xu H, Denny JC, *et al*. The emerging role of electronic medical records in pharmacogenomics. *Clin Pharmacol Ther* 2011;**89**:379–86.
3. **Wilke RA,** Berg RL, Peissig P, *et al*. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. *Clin Med Res* 2007;**5**:1–7.
4. **Wilke RA,** Berg RL, Linneman JG, *et al*. Characterization of low-density lipoprotein cholesterol-lowering efficacy for atorvastatin in a population-based DNA biorepository. *Basic Clin Pharmacol Toxicol* 2008;**103**:354–9.
5. **McCarty CA,** Chisholm RL, Chute CG, *et al*. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011;**4**:13.
6. **Liao KP,** Cai T, Gainer V, *et al*. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)* 2010;**62**:1120–7.
7. **Bourgeois FC,** Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. *Arch Intern Med* 2010;**170**:1989–95.
8. **Smith PC,** Araya-Guerra R, Bublitz C, *et al*. Missing clinical information during primary care visits. *JAMA* 2005;**293**:565–71.
9. **Elder NC,** Hickner J. Missing clinical information: the system is down. *JAMA* 2005;**293**:617–19.
10. **Elder NC,** Vonder Meulen M, Cassedy A. The identification of medical errors by family physicians during outpatient visits. *Ann Fam Med* 2004;**2**:125–9.
11. **Cwinn MA,** Forster AJ, Cwinn AA, *et al*. Prevalence of information gaps for seniors transferred from nursing homes to the emergency department. *CJEM* 2009;**11**:462–71.
12. **Kim J,** Chuun D, Shah A, *et al*. Prevalence and impact of information gaps in the emergency department. *AMIA Annu Symp Proc* 2008:866.
13. **Stiell A,** Forster AJ, Stiell IG, *et al*. Prevalence of information gaps in the emergency department and the effect on patient outcomes. *CMAJ* 2003;**169**:1023–8.
14. **Cox A,** Rutter M, Yule B, *et al*. Bias resulting from missing information: some epidemiological findings. *Br J Prev Soc Med* 1977;**31**:131–6.
15. **Saaddine JB,** Engelgau MM, Beckles GL, *et al*. A diabetes report card for the United States: quality of care in the 1990s. *Ann Intern Med* 2002;**136**:565–74.

16. **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;**19**:212—18.
17. **Kashner TM.** Agreement between administrative files and written medical records: a case of the Department of Veterans Affairs. *Med Care* 1998;**36**:1324—36.
18. **Hebert PL,** Geiss LS, Tierney EF, *et al*. Identifying persons with diabetes using medicare claims data. *Am J Med Qual* 1999;**14**:270—7.
19. **St Sauver JL,** Grossardt BR, Yawn BP, *et al*. Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project. *Am J Epidemiol* 2011;**173**:1059—68.
20. **Melton LJ 3rd Jr.** History of the Rochester epidemiology project. *Mayo Clin Proc* 1996;**71**:266—74.
21. **Melton LJ 3rd.** The threat to medical-records research. *N Engl J Med* 1997;**337**:1466—70.
22. **McNemar Q.** Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947;**12**:153—7.
23. **Team RDC.** *R: A language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2011.
24. *National Diabetes Statistics*. 2011. http://diabetes.niddk.nih.gov/dm/pubs/statistics/
25. *Wisconsin Diabetes Mellitus Essential Care Guidelines*. 2011. http://www.dhs.wisconsin.gov/health/diabetes/PDFs/GL01.pdf
26. **Martinez M,** Khlat M, Leboyer M, *et al*. Performance of linkage analysis under misclassification error when the genetic model is unknown. *Genet Epidemiol* 1989;**6**:253—8.
27. *Beacon*. http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__onc_beacon_community_program__improving_health_through_health_it/1805
28. **Tan PN,** Steinbach M, Kumar V. *Introduction to Data Mining*. Boston: Pearson Addison Wesley, 2006.
29. **DeLisle S,** South B, Anthony JA, *et al*. Combining free text and structured electronic medical record entries to detect acute respiratory infections. *PLoS One* 2010;**5**:e13377.
30. **Li L,** Chase HS, Patel CO, *et al*. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. *AMIA Annu Symp Proc* 2008:404—8.
31. **Wei W,** Tao C, Jiang G, *et al*. A high throughput semantic concept frequency based approach for patient identification: a case study using type 2 diabetes mellitus clinical notes. *AMIA Annu Symp Proc* 2010:857—61.
32. **Turchin A,** Kohane IS, Pendergrass ML. Identification of patients with diabetes from the text of physician notes in the electronic medical record. *Diabetes Care* 2005;**28**:1794—5.
33. **O'Malley KJ,** Cook KF, Price MD, *et al*. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005;**40**:1620—39.
34. **Leibson CL,** O'Brien PC, Atkinson E, *et al*. Relative contributions of incidence and survival to increasing prevalence of adult-onset diabetes mellitus: a population-based study. *Am J Epidemiol* 1997;**146**:12—22.