Published in final edited form as:
Int J Med Inform. 2013 April ; 82(4): 239–247. doi:10.1016/j.ijmedinf.2012.05.015.

The Absence of Longitudinal Data Limits the Accuracy of High-Throughput Clinical Phenotyping for Identifying Type 2 Diabetes Mellitus Subjects

Wei-Qi Wei, MM[1,2], **Cynthia L. Leibson, PhD**[3], **Jeanine E. Ransom**[2], **Abel N. Kho, MD, MS**[4], and **Christopher G. Chute, MD, DrPH**[2]

[1]Institute for Health Informatics, University of Minnesota, Twin Cities, MN

[2]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN

[3]Division of Epidemiology, Mayo Clinic, Rochester, MN

[4]Divisions of General Internal Medicine and Biomedical Informatics, Northwestern University Feinberg School of Medicine, Chicago, IL

Abstract

Purpose—To evaluate the impact of insufficient longitudinal data on the accuracy of a high-throughput clinical phenotyping (HTCP) algorithm for identifying 1) patients with type 2 diabetes mellitus (T2DM) and 2) patients with no diabetes.

Methods—Retrospective study conducted at Mayo Clinic in Rochester, Minnesota. Eligible subjects were Olmsted County residents with ≥1 Mayo Clinic encounter in each of three time periods : 1) 2007, 2) from 1997 through 2006, and 3) before 1997 (N= 54,283). Diabetes relevant electronic medical record (EMR) data about diagnoses, laboratories, and medications were used. We employed the HTCP algorithm to categorize individuals as T2DM cases and non-diabetes controls. Considering the full 11 years (1997–2007) as the gold standard, we compared gold-standard categorizations with those using data for 10 subsequent intervals, ranging from 1998–2007 (10-year data) to 2007 (1-year data). Positive predictive values (PPVs) and false-negative rates (FNRs) were calculated. McNemar tests were used to determine whether categorizations using shorter time periods differed from the gold standard. Statistical significance was defined as $P<.05$.

Results—We identified 2,770 T2DM cases and 21,005 controls when the algorithm was applied using 11-year data. Using 2007 data alone, PPVs and FNRs respectively were 70% and 25% for

Corresponding author: Christopher G. Chute, Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 First St SW, Rochester, MN 55905 (chute@mayo.edu). Telephone: (507) 284-5541.

Authors' contributions

Study concept and design: Wei-Qi Wei, Cynthia L. Leibson, Abel N. Kho, and Christopher G. Chute.
Acquisition of data: Wei-Qi Wei and Jeanine E. Ransom.
Analysis and interpretation of data: Wei-Qi Wei, Cynthia L. Leibson, and Jeanine E. Ransom.
Drafting of the manuscript: Wei-Qi Wei, Cynthia L. Leibson, Jeanine E. Ransom, Abel N. Kho, and Christopher G. Chute.

Conflict of interest

There are no conflicts of interests.

case identification and 59% and 67% for control identification. All time frames differed significantly from the gold standard, except for the 10-year period.

Conclusions—The accuracy of the algorithm reduced remarkably as data were limited to shorter observation periods. This impact should be considered carefully when designing/executing HTCP algorithms.

Keywords

diabetes mellitus; electronic medical record; phenotype; data aggregation; medical informatics; research subject selection

Introduction

Clinically relevant genomic studies offer the hope of improving routing care by discovering knowledge associated between genetic variants and disease. In order to derive statistically powerful conclusions, a considerable number of subjects are often required on the basis of disease, symptoms, or related findings. This subject selection process usually consumes substantial time and human efforts to gather, abstract, and review patients' medical records. Cost-effective ways are demanded to perform clinical phenotyping within large populations [1].

Recently, the increased adoption of electronic medical record (EMR) systems has provided a potential tool to reduce the inefficiencies of manual medical record review [2, 3]. By leveraging machine-processable content through an EMR system, clinical researchers can develop a high-throughput clinical phenotyping (HTCP) algorithm (a set of EMR-based inclusion and exclusion subject selection criteria), execute the algorithm against an existing EMR system, and rapidly obtain a large pool of potentially eligible study subjects [4–6].

The Electronic Medical Records and Genomics (eMERGE) Network, a national consortium funded by the National Human Genome Research Institute, has devoted substantial efforts to exploring the possibility of leveraging EMRs for HTCP [7]. The eMERGE I Network consisted of 5 leading medical centers in the United States: Mayo Clinic, Rochester, Minnesota; Northwestern University Medical Center, Chicago, Illinois; Vanderbilt University Medical Center, Nashville, Tennessee; Marshfield Clinic, Marshfield, Wisconsin; and Group Health Cooperative in collaboration with University of Washington, Seattle, Washington. Thirteen robust HTCP algorithms had been created by the end of May 2011. All HTCP algorithms were validated across the 5 participating centers to ensure that each of them was transportable and that various institutions can execute it to efficiently and thereby accurately identify subjects eligible for clinical research.

However, the role of longitudinal data on the accuracy of HTCP algorithms continues to be a concern. Insufficient longitudinal data occurs when a patient's EMR data is limited to a short time frame, e.g., the patient's data were collected before the EMR system was implemented at a center or the patient was seen at that center only for a short period. In either instance, only a certain number of years of EMR data, instead of a patient's complete longitudinal data, would be available when executing the algorithm. The unavailable longitudinal data may be crucial to qualify or disqualify study subjects. For example, distinguishing persons with diabetes from persons with no diabetes, glucose values are frequently a key factor used in this distinction. But if the patient is well controlled off medication, he/she may not have abnormal glucose values available in the EMR data. The lack of these data may bring about subject selection errors, lead to sampling bias, and, more importantly, risk misleading results of following studies[8]. The disadvantages of short periods of observation for describing patient characteristics have been reported in several

epidemiological studies by our group and others, e.g., distinguishing incident from prevalent cases and identifying risk factors [9–14].

However, the impact of insufficient longitudinal data on an HTCP algorithm has not been explicitly investigated. This present study aimed to provide a novel demonstration of this impact of insufficient longitudinal data on an HTCP algorithm developed within the eMERGE Network for specifying subjects with T2DM. We limited the data from one single medical center (Mayo Clinic) alone because 1) the majority of medical care received by Olmsted County residents is provided by Mayo Clinic and the EMR system at Mayo Clinic has been implemented over a decade [8], therefore, the EMR system at Mayo Clinic has rich longitudinal EMR data available for answering the proposed question; and 2) using the EMR data at one single medical center isolates the temporal issue from other variables, e.g. spatial data fragmentation, which has been addressed by other studies [14].

The eMERGE T2DM algorithm

T2DM accounts for substantial morbidity and mortality from adverse effects on cardiovascular risk and disease-specific complications such as blindness and renal failure [9]. The increasing global prevalence of T2DM has imposed an enormous public health burden [10]. However, the disease is a poorly understood [11]. The exploring knowledge of associated knowledge between genetic variants and T2DM will deliver clues to the processes involved in disease pathogenesis, offer potential targets for new drugs, and, hopefully, lead to a cure [11, 12]. The eMERGE T2DM algorithm is such an EMR based effort aiming to quickly and precisely identifying subjects with T2DM for genotype and phenotype associated analyses [13].

As described in detail elsewhere [13, 14], the eMERGE T2DM algorithm was developed by researchers from Northwestern University and enhanced by other participating institutes in the eMERGE Network. The primary goal of developing this algorithm was to precisely identify subjects who truly have T2DM rather than to identify all subjects with T2DM within the population. In other words, the algorithm was not designed to perform a dichotomous classification, i.e. subjects with T2DM or subjects without T2DM, but to achieve a high positive predictive value (PPV), i.e. precision, of identifying patients with T2DM, or T2DM cases, and to avoid confounding by including individuals without any type of diabetes mellitus (DM) or those individuals with T1DM. Similarly, with respect to unaffected controls or non-DM controls, the goal of the algorithm was to maximize the PPV of identifying individuals without DM of any type, excluding even those individuals at risk for DM that had not yet manifest (i.e., pre-DM).

Previous evidence suggests that using ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) codes alone is problematic for identifying diabetes subjects generally [15, 16]. More importantly, T2DM cases identified by using ICD-9-CM codes alone could be contaminated with T1DM cases because many patients are assigned the code for “diabetes mellitus, unspecified type” and some patients with T2DM diagnosis codes are actually T1DM individuals who have been wrongly assigned a code for T2DM. To avoid the limitations of relying on ICD diagnosis codes alone, the algorithm developers supplemented the use of diagnosis codes with relevant medication prescriptions and associated laboratory results (Figure 1 and Figure 2). For high precision purpose, the algorithm used different diagnostic lab cutoffs from clinically recommended. Namely, the algorithm used a combination of multiple EMR data sources and adopted a different set of criteria than normally used to diagnose T2DM.

A previous validation of the algorithm on a few randomly chosen cases demonstrated its precision is comparable with clinicians’ medical record review [13]. Importantly, however,

the time frame available for review was the same as the EMR data available for use. The role of insufficient longitudinal data and its impact on the accuracy of the algorithm was not addressed. This present study aimed to evaluate the impact of insufficient longitudinal data on the eMERGE T2DM algorithm, to determinate the absence of which data (diagnostic codes, laboratory tests, or medications) contributes to subject misclassification, and to suggest how best to leverage EMR to optimize subject selection. We chose to evaluate the impact on the basis of the eMERGE T2DM algorithm because it involves virtually all structured EMR data (i.e. diagnosis, laboratory values, and medication) and its accuracy within a single medical center has been evaluated.

Methods

Study Setting

This retrospective cohort study was conducted in Olmsted County, Minnesota (2010 census = 144,248). Rochester, the county seat is geographically isolated (80 miles from any other urban center), and home to Mayo Clinic, one of the world's largest medical centers. Each year, more than half of the County population is examined at Mayo Clinic, and the vast majority of residents have at least one Mayo Clinic encounter during any 3-year period [8]. Since 1907, every patient seen at Mayo Clinic has been assigned a unique identifier, and data from every encounter are contained in a patient-based medical record [17], thus individuals can be followed across settings and over time. Mayo Clinic's EMR system began in 1993. Today, Mayo Clinic has one of the most advanced EMR systems in the United States [18], providing a unique opportunity to investigate the value of longitudinal EMR data for subject selection in HTCP.

Eligible Subjects

The present study was approved by the Mayo Clinic Institutional Review Board. Patients were excluded who refused authorization for use of their medical records at Mayo Clinic in research, typically less than 5% [19]. To ensure that subjects had more than 11 years of EMR data at Mayo, all eligible subjects had to have been Olmsted County residents with 1 Mayo Clinic encounter in each of three time periods: 1) 2007, 2) from 1997 through 2006, and 3) before 1997 (earliest date for which Mayo's data were fully implemented).

EMR Data

We used data available through the EMR system at Mayo Clinic. Administrative claims data were searched to determine the presence or absence of diabetes-related ICD-9-CM codes. Outpatient laboratory data were used to determine whether or not a patient had abnormal laboratory glucose or HbA1c values. For diabetes medication details, we used data from Mayo Clinic's outpatient prescription database. We manually checked the medication data and produced a list of generic drug names, brand names, synonyms, and abbreviations for diabetes-relevant medications (see appendix). To determine whether or not a patient had been prescribed any such medications, we searched the outpatient prescription database for the terms on the list.

Gold Standard

We hypothesized that the more complete data over extensive periods of observation, the more accurate the result of the algorithm. We applied the eMERGE T2DM algorithm to each eligible patient's 11 years (1997–2007) of EMR data. The categorization of individuals as T2DM cases and non-DM controls when using these data was considered the gold standard in this study.

Study Design

Results obtained using the gold standard were compared with results from executing the eMERGE T2DM algorithm under 10 time separate frames: 1 year (2007), 2 years (2006–2007), 3 years (2005–2007), 4 years (2004–2007), 5 years (2003–2007), 6 years (2002–2007), 7 years (2001–2007), 8 years (2000–2007), 9 years (1999–2007), and 10 years (1998–2007).

Statistical Analysis

We calculated PPVs and false-negative rates (FNRs) to evaluate the misclassification errors that resulted from shorter time periods of data. The PPV is the true-positive rate (determined by the gold standard, i.e., using 11 years of EMR data) of subjects positively identified when less longitudinal data were used, defined as the following:

$$PPV = \frac{\text{No. of True–Positive Cases (True–Positive Controls)}}{\text{No. of True–Positive Cases (True–Positive Controls)} + \text{No. of False–Positive Cases (No. of False–Positive Controls)}}$$

The FNR is the false-negative rate of subjects positively identified with the gold standard, calculated with the following:

$$\text{FNR} = \frac{\text{No. of False–Negative Cases (False–Negative Controls)}}{\text{No. of True–Positive Cases (True–Positive Controls)} + \text{No. of False–Negative Cases (False–Negative Controls)}}$$

McNemar test[20] has been widely used for testing whether the row and column marginal frequencies of a 2×2 table are homogeneous. We used this method to determine whether the categorization using less longitudinal data differed from the gold standard. All tests were followed by the Bonferroni procedure to correct for multiple comparisons. Statistical significance was accepted when adjusted P was < .05. All P values reported were original P values. All data are presented as mean and standard deviation (SD). Statistical analysis was performed with R for Windows (version 2.11.1) [21].

Results

Among the 139,654 Olmsted County residents in 2007, a total number of 86,294 had at least 1 Mayo Clinic visit in 2007, of whom 54,283 had both >= 1 Mayo Clinic visit from 01/01/1997 through 12/31/2006, and >=1 Mayo Clinic visit before 1997. These 54,283 patients were eligible for the study (mean [SD] age, 46.6 [21.3] years; 56% female).

Case Identification

From the 54,283 eligible patients, 2,770 T2DM cases (mean [SD] age, 64.6 [14.1] years; 47% female) were identified when 11 years of EMR data were used. These 2,770 T2DM cases were considered true, i.e., gold standard, T2DM.

T2DM case identification errors increased as the time frame of available EMR data decreased (Table 1). Errors included patients with T1DM or patients with no DM who were misclassified as T2DM (false-positive) or true T2DM cases who were misclassified as T1DM or no DM (false-negative).

When 10 years of EMR data were used—only 1 year less than the gold standard—13 of 2,768 identified T2DM cases were false-positive (PPV, 99.5% [2,755/2,768]), and 15 of 2,770 true T2DM cases were incorrectly excluded (FNR, 0.5%). When 5 years of EMR data

were used (2003–2007), 435 of 3,051 identified T2DM cases were false-positive (PPV, 86% [2,616/3,051]), and 154 of 2,770 true T2DM cases were incorrectly excluded (FNR, 6%). When only 1 year of EMR data (2007) was used, 881 of 2,970 identified T2DM cases were false-positive and 681 of 2,770 true T2DM cases were incorrectly excluded. The PPV of the algorithm for T2DM case identification decreased to 70% (2,089/2,970) and the FNR increased to 25% (681/2,770). McNemar tests indicated that a significant difference existed between the gold standard and the categorizations when 9 years or less of EMR data were used (Table 1).

With respect to which eMERGE inclusion/exclusion criteria accounted for the misclassification using shorter time frames (see Figure 1), all false-positive T2DM cases resulted from missing T1DM diagnosis (Table 2). The majority of true T2DM cases incorrectly excluded using shorter time frames resulted from diagnosis codes for T2DM or prior abnormal laboratory reports (Table 3) that were exclude due to shorter time frames.

Control Identification

From the 54,283 eligible patients, 21,005 non-DM controls (mean [SD] age, 40.7 [17.8] years; 62% female) were identified when 11 years of EMR data were used. These subjects were considered true non-DM controls of the gold standard.

As with T2DM case identification, non-DM control identification errors increased as the time frame of available EMR data decreased (Table 1). When 10 years of EMR data were used, 283 of 20,807 identified non-DM controls were false-positive (PPV, 99% [20,524/20,807]) and 481 of 21,005 true-positive non-DM controls were incorrectly excluded; FNR, 2%). When 5 years of EMR data were used, the PPV was 84% (16,584/19,759), and the FNR was 21% (4,421/21,005). When only 1 year of EMR data was used, 4,690 of 11,576 identified non-DM controls were false-positive and 14,119 of 21,005 true non-DM controls were incorrectly excluded. The PPV of the algorithm for non-DM control identification decreased to 59% (6,886/11,576), and the FNR increased to 67%. McNemar tests indicated that a significant difference existed between the gold standard and the categorizations when 10 years of EMR data or less were used (Table 1).

With respect to which eMERGE inclusion/exclusion criteria accounted for the misclassification of non-DM controls using shorter time frames (see Figure 2), approximately 99% of false-positive non-DM controls resulted from missing diabetes diagnosis codes or missing prior abnormal laboratory reports (Table 2). The majority of false-negative non-DM controls resulted from missing laboratory reports (Table 3). The remaining false-negative non-DM controls were missed because they did not have the required number of face-to-face encounters required by the algorithm within the time frame under consideration.

Discussion

Subject selection has become a tedious obstacle to conducting more clinical phenotype related research in detail. As the influence of medical informatics grows and EMR systems —the core application of medical informatics—expand, HTCP, through leveraging the machine-processable clinical data, will have a pivotal role in optimizing this inefficient process [6, 7].

However, few medical centers today have a patient’s complete longitudinal data in their EMR systems. Therefore, although a patient’s medical record is a long-term individual history, it can be available for clinical research cross-sectionally within only varying time frames. A demonstration of the impact of the insufficient longitudinal data on the accuracy

of an HTCP algorithm is critical for understanding how best to use EMR to improve subject selection. However, this demonstration had not yet been done, possibly because of the difficulty in obtaining an EMR resource that maintains relatively complete longitudinal data. By taking advantage of the richness of data in the EMR system at Mayo Clinic, we accomplished such a novel investigation.

Our results, combined with findings from other studies[22–27], show the advantage of access to more data over longer time periods. The T2DM HTCP algorithm has been well designed and carefully validated within the eMERGE Network. Its precision has been demonstrated comparable with clinician review [13]. Nevertheless, our findings suggest that the absent longitudinal data significantly change the subject categorization of the algorithm. For both T2DM case and non-DM control identifications, statistical analyses indicated that a significant difference between the categorization of individuals when 11 years of EMR data were used compared to when less longitudinal data are available. The differences between categorizations results from not only a large proportion of false-negatives, but also a considerable number of false-positives. The absence of historical diagnosis data, prior DM-related laboratory results, or medication use history contributes to the misclassifications. Specifically, the absent historical diagnosis data contributes to the majority of false-positive T2DM cases. The absent prior laboratory results lead to the most false-negative non-DM controls.

Both PPV and FNR notably changed when the time frame of EMR data was reduced. Even though the algorithm was designed primarily to achieve a robust and high PPV, the PPV dramatically decreased to below 80% when less than 4 years of EMR data were available. When only a couple of year of EMR data were available, the PPV of the case identification decreased to 70%, which is not sufficient for most genotype-and-phenotype-associated analyses. The FNRs rose as well when longitudinal data were insufficient, especially for the control identification. As hypothesized, PPV and FNR are improved when data can be obtained over extensive periods of observation compared with data limited to a short period.

Even though a 100% PPV or 0% FNR may be an unachievable goal for an EMR-based algorithm, a poor PPV or FNR could result in sampling bias and risk serious distortions in the results of following studies [28]. Our results suggested that ≥7 years of EMR data should be used in order to achieve a >90% PPV. Unstructured EMR data, e.g. clinical notes, may have some relevant description of phenotypes that are unavailable in structured EMR data. Incorporating unstructured EMR data into subject selection criteria may relieve the problem caused by the unavailability of longitudinal data, e.g. problem lists can be used with ICD codes to determinate whether or not a patient has a disease. Our previous work, along with other studies, has shown the potential of unstructured EMR data to be used for subject selection tasks [4, 29–32].

Researchers who execute an HTCP algorithm on their EMR systems should be aware that it is error prone when EMR data are available from only a limited time. However, obtaining EMR data over longer periods helps only when the additional longitudinal data are relevant for the condition under study. Chronic conditions, such as T2DM in this study, are more appropriate than emergence conditions, such as fractures or acute infectious diseases. In addition, if a patient is seen by multiple providers, a single medical center may not have a patient's complete medical history. In that situation, the accuracy of an algorithm will not improve even when long-term clinical data are available.

The present study has several limitations. One is the gold standard. Because of unavoidable random or systematic errors, such as physician experience, communication quality between a patient and a clinician, and coding quality, it is extremely difficult to obtain a patient's

actual condition [33, 34]. We hypothesize that the more longitudinal data obtained over extensive periods of observation, the closer the categorization of a patient to the actual condition. Mayo Clinic maintains a century of diagnostic coded data. However, reliable EMR medication data are only available since 1997. We used all 11 years of reliable EMR data to create the gold standard in our study. Our present data does not identify a point of diminishing returns.

Our investigation was limited to a single geographical region, which is predominantly white. Compared to US whites, the age- and sex -distribution is similar; however median income and education levels are higher [8]. No single geographic area is representative of all others; however, the under-representation of minorities and the limitation to a single medical center compromises the generalizability of findings to other racial/ethnic groups and different health care environments. Because the number of providers from which Olmsted County residents received care is limited, subject selection errors due to insufficient longitudinal data that we demonstrated in this study are, most likely, an underestimate of that occurring in other centers.

HTCP is an important secondary and meaningful use of EMR application. We demonstrated the impact of the insufficient longitudinal data on the basis of one algorithm. For a more complete evaluation of the impact, this study should be repeated with a broader spectrum of HTCP algorithms.

Conclusion and Suggestion

The present study provided a previously unavailable demonstration of the impact of insufficient longitudinal data on the accuracy of HTCP. Our results showed marked changes in the PPVs and the FNRs of the algorithm for identifying both cases and controls, depending on the completeness of longitudinal data available for each patient. The impact due to absent longitudinal data should be carefully considered when designing an HTCP algorithm or executing one.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the Biomedical Informatics and Computational Biology Graduate Traineeship Program, University of Minnesota; the eMERGE project, NIH U01 HG04599; and the Strategic Health IT Advanced Research Projects program, #90TR0002-01Z-02. We wish to acknowledge fruitful discussions with Dr. High Seng Chai and Dr. Pedro Caraballo.

Abbreviations

DM	diabetes mellitus
eMERGE	Electronic Medical Records and Genomics
EMR	electronic medical record
FNR	false-negative rate
HTCP	high-throughput clinical phenotyping
ICD-9-CM	International Classification of Diseases-9-Clinical Modification
PPV	positive predictive value

SD	Standard Deviation
T1DM	type 1 diabetes mellitus
T2DM	type 2 diabetes mellitus

References

1. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011; 12(6):417–28. [PubMed: 21587298]
2. Ritchie MD, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet. 2010; 86(4):560–72. [PubMed: 20362271]
3. Wilke RA, et al. The emerging role of electronic medical records in pharmacogenomics. Clin Pharmacol Ther. 2011; 89(3):379–86. [PubMed: 21248726]
4. Liao KP, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken). 62(8):1120–7. [PubMed: 20235204]
5. Wilke RA, et al. Characterization of low-density lipoprotein cholesterol-lowering efficacy for atorvastatin in a population-based DNA biorepository. Basic Clin Pharmacol Toxicol. 2008; 103(4): 354–9. [PubMed: 18834356]
6. Wilke RA, et al. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. Clin Med Res. 2007; 5(1):1–7. [PubMed: 17456828]
7. McCarty CA, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 4:13. [PubMed: 21269473]
8. Melton LJ 3rd. History of the Rochester Epidemiology Project. Mayo Clin Proc. 1996; 71(3):266–74. [PubMed: 8594285]
9. Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. Nature. 2001; 414(6865):782–7. [PubMed: 11742409]
10. Saaddine JB, et al. A diabetes report card for the United States: quality of care in the 1990s. Ann Intern Med. 2002; 136(8):565–74. [PubMed: 11955024]
11. McCarthy MI. Genomics, type 2 diabetes, and obesity. N Engl J Med. 2010; 363(24):2339–50. [PubMed: 21142536]
12. Travers ME, McCarthy MI. Type 2 diabetes and obesity: genomics and the clinic. Hum Genet. 2011; 130(1):41–58. [PubMed: 21647602]
13. Kho AN, 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(2):212–8. [PubMed: 22101970]
14. Wei WQ, 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. [PubMed: 22249968]
15. Kashner TM. Agreement between administrative files and written medical records: a case of the Department of Veterans Affairs. Med Care. 1998; 36(9):1324–36. [PubMed: 9749656]
16. Hebert PL, et al. Identifying persons with diabetes using Medicare claims data. Am J Med Qual. 1999; 14(6):270–7. [PubMed: 10624032]
17. Kurland LT, Molgaard CA. The patient record in epidemiology. Sci Am. 1981; 245(4):54–63. [PubMed: 7027437]
18. Chute CG, et al. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc. 17(2):131–5. [PubMed: 20190054]
19. Melton LJ 3rd. The threat to medical-records research. N Engl J Med. 1997; 337(20):1466–70. [PubMed: 9380105]
20. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947; 12(2):153–7. [PubMed: 20254758]

21. Team RDC. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
22. Brown AW, et al. Long-term survival after traumatic brain injury: a population-based analysis. NeuroRehabilitation. 2004; 19(1):37–43. [PubMed: 14988586]
23. Leibson CL, et al. Comorbid conditions associated with Parkinson's disease: a population-based study. Mov Disord. 2006; 21(4):446–55. [PubMed: 16161155]
24. Leibson CL, 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(1):12–22. [PubMed: 9215219]
25. Pannala R, et al. Temporal association of changes in fasting blood glucose and body mass index with diagnosis of pancreatic cancer. Am J Gastroenterol. 2009; 104(9):2318–25. [PubMed: 19513024]
26. Rocca WA, et al. Long-term effects of bilateral oophorectomy on brain aging: unanswered questions from the Mayo Clinic Cohort Study of Oophorectomy and Aging. Womens Health (Lond Engl). 2009; 5(1):39–48. [PubMed: 19102639]
27. Wannamethee SG, et al. Impact of diabetes on cardiovascular disease risk and all-cause mortality in older men: influence of age at onset, diabetes duration, and established and novel risk factors. Arch Intern Med. 171(5):404–10. [PubMed: 21403036]
28. Martinez M, et al. Performance of linkage analysis under misclassification error when the genetic model is unknown. Genet Epidemiol. 1989; 6(1):253–8. [PubMed: 2731714]
29. DeLisle S, et al. Combining free text and structured electronic medical record entries to detect acute respiratory infections. PLoS One. 5(10):e13377. [PubMed: 20976281]
30. Li L, 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. [PubMed: 18999285]
31. Wei, W., et al. A High Throughput Semantic Concept Frequency Based Approach for Patient Identification: A Case Study Using Type 2 Diabetes Mellitus Clinical Notes. Proc AMIA Symp; 2010.
32. Turchin A I, Kohane S, Pendergrass ML. Identification of patients with diabetes from the text of physician notes in the electronic medical record. Diabetes Care. 2005; 28(7):1794–5. [PubMed: 15983338]
33. O'Malley KJ, et al. Measuring diagnoses: ICD code accuracy. Health Serv Res. 2005; 40(5 Pt 2): 1620–39. [PubMed: 16178999]
34. Spolaore P, et al. Measuring accuracy of discharge diagnoses for a region-wide surveillance of hospitalized strokes. Stroke. 2005; 36(5):1031–4. [PubMed: 15790948]

Summary Table

What Was Already Known on the Topic?

- Clinical research requires identifying cohorts of potentially eligible subjects on the basis of disease, symptoms, or related findings. This process usually consumes substantial time and human efforts to gather, abstract, and review patients' charts.
- HTCP leverages machine-processable EMR data, improving the inefficiency of this subject selection process.
- Many disadvantages of the inability to access clinical data over longer periods have been demonstrated previously. But little is known of the impact of insufficient longitudinal data on the accuracy of HTCP.

What This Study Added to Our Knowledge?

- Our study demonstrated that insufficient longitudinal data reduced the accuracy of HTCP for identifying both cases and controls, depending on the completeness of longitudinal data available for each patient.
- The impact due to insufficient longitudinal data should be carefully considered when designing an HTCP algorithm or executing one.

Highlights

- electronic medical record
- phenotype; data aggregation
- medical informatics
- research subject selection
- diabetes mellitus

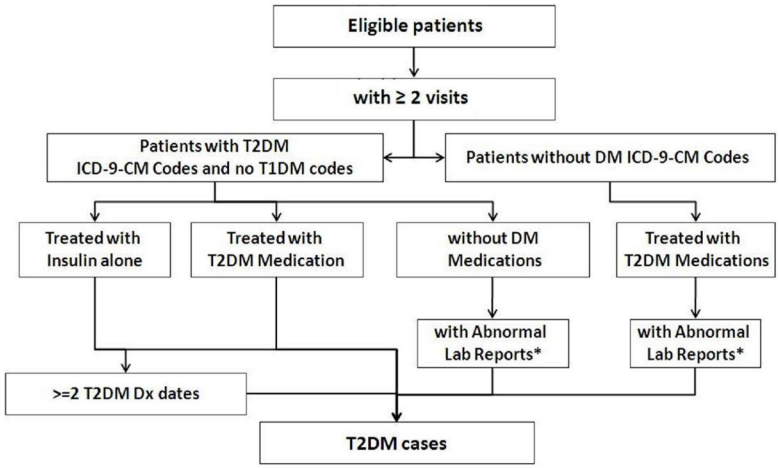

Figure 1.

The eMERGE Algorithm for Identifying T2DM Cases

*Random glucose > 200 mg/dl, Fasting glucose > 125 mg/dl, hemoglobin A1c ≥ 6.5%

Abbreviations: DM, diabetes mellitus; Dx, diagnosis; eMERGE, Electronic Medical Records and Genomics; HbA_{1c}, hemoglobin A_{1c}; ICD-9-CM, International Classification of Diseases, 9th Revision, Clinical Modification; Rx, prescription; T2DM, type 2 diabetes mellitus; T1DM, type 1 diabetes mellitus.

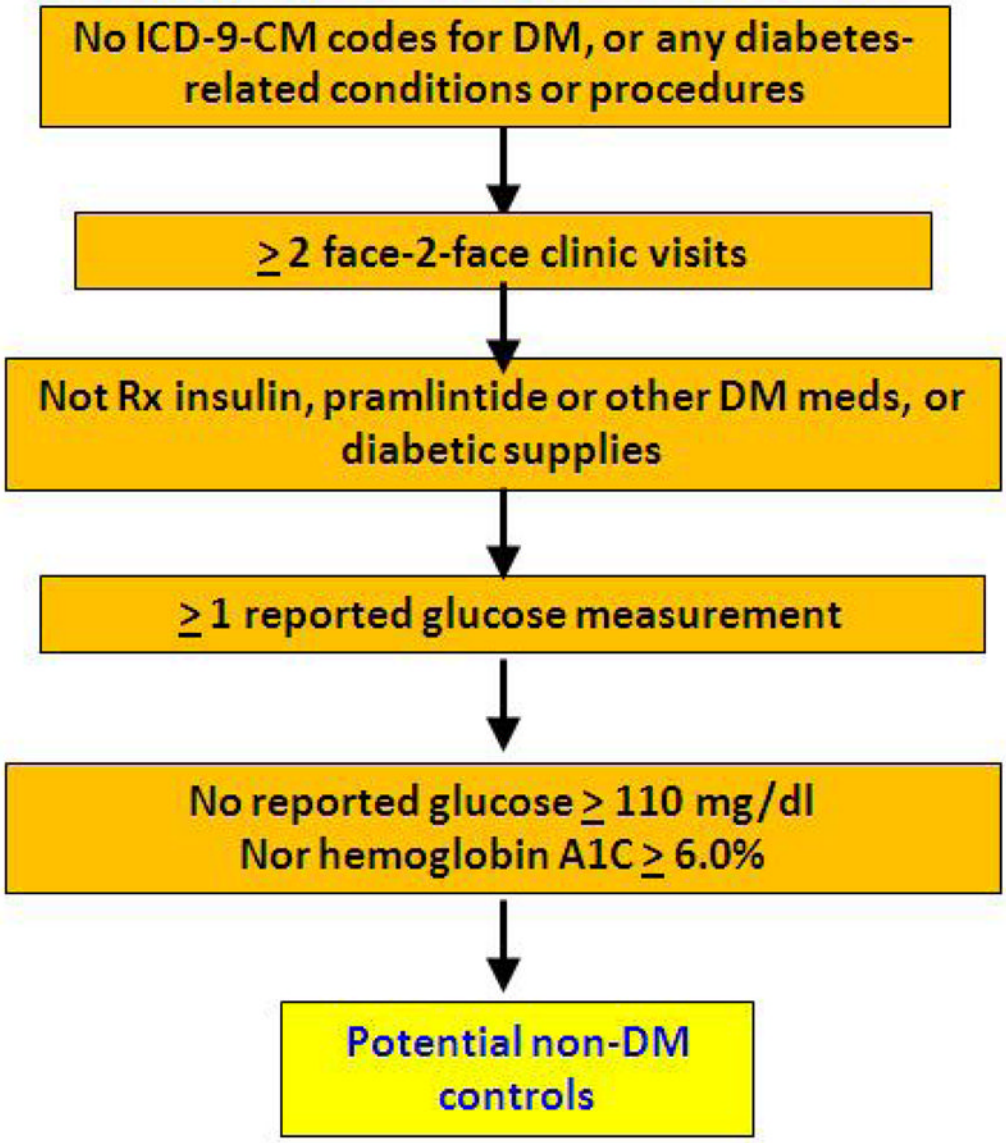

Figure 2.
The eMERGE Algorithm for Identifying non-DM Controls
Abbreviations: DM, diabetes mellitus; Dx, diagnosis; eMERGE, Electronic Medical Records and Genomics; HbA_{1c}, hemoglobin A_{1c}; ICD-9-CM, International Classification of Diseases, 9^{th} Revision, Clinical Modification; Rx, prescription; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus.

Table 1

The Outputs of Case and Control Identification When Applying the Algorithm on EMR Data within Different Time Frames.

	Time Frame of EMR Data	Patients with ≥ 2 visits	Identified Subjects (TP+FP), No.	TPs, No.	FPs, No.	TNs, No.	FNs, No.	PPV (TP/(TP+FP)), %	FNR (FN/(TP+FN)), %	*P* Value
Case	2007 (1 yr)	74212	2970	2089	881	50632	681	70%	25%	*P*<0.01
	2006–2007 (2 yrs)	82679	3280	2366	914	50599	404	72%	15%	*P*<0.01
	2005–2007 (3 yrs)	83792	3374	2466	908	50605	304	73%	11%	*P*<0.01
	2004–2007 (4 yrs)	84326	3273	2550	723	50790	220	78%	8%	*P*<0.01
	2003–2007 (5 yrs)	84617	3051	2616	435	51078	154	86%	6%	*P*<0.01
	2002–2007 (6 yrs)	84788	2967	2650	317	51196	120	89%	4%	*P*<0.01
	2001–2007 (7 yrs)	84903	2936	2692	244	51269	78	92%	3%	*P*<0.01
	2000–2007 (8 yrs)	84993	2919	2721	198	51315	49	93%	2%	*P*<0.01
	1999–2007 (9 yrs)	85072	2858	2743	115	51398	27	96%	1%	*P*<0.01
	1998–2007 (10 yrs)	85125	2768	2755	13	51500	15	99.5%	0.5%	*P*=0.85
	1997–2007 (gold standard)	*85172*	*2770*	*2770*	*0*	*51513*	*0*	*100%*	*0%*	*NA*
Control	2007 (1 yr)	74212	11576	6886	4690	28588	14119	59%	67%	*P*<0.01
	2006–2007 (2 yrs)	82679	15700	10948	4752	28526	10057	70%	48%	*P*<0.01
	2005–2007 (3 yrs)	83792	17524	13352	4172	29106	7653	76%	36%	*P*<0.01
	2004–2007 (4 yrs)	84326	18800	15084	3716	29562	5921	80%	28%	*P*<0.01
	2003–2007 (5 yrs)	84617	19759	16584	3175	30103	4421	84%	21%	*P*<0.01
	2002–2007 (6 yrs)	84788	20326	17688	2638	30640	3317	87%	16%	*P*<0.01
	2001–2007 (7 yrs)	84903	20080	18606	1474	31804	2399	93%	11%	*P*<0.01
	2000–2007 (8 yrs)	84993	20297	19364	933	32345	1641	95%	8%	*P*<0.01
	1999–2007 (9 yrs)	85072	20537	19968	569	32709	1037	97%	5%	*P*<0.01
	1998–2007 (10 yrs)	85125	20807	20524	283	32995	481	99%	2%	*P*<0.01
	1997–2007 (gold standard)	*85172*	*21005*	*21005*	*0*	*33278*	*0*	*100%*	*0%*	*NA*

Abbreviations: EMR, electronic medical record; FNR, false-negative rate; FN, false-negative; FP, false-positive; NA, not applicable; PPV, positive predictive value; TN, true-negative; TP, true-positive.

Table 2

The Numbers and the Reasons that Contribute to False-Positive Subjects.

	Time Frame of EMR Data	FP Subjects, No.	Missing T1DM Diagnosis Codes, No. (%)
Case	2007 (1 yr)	881	881 (100)
	2006–2007 (2 yrs)	914	914 (100)
	2005–2007 (3 yrs)	908	908 (100)
	2004–2007 (4 yrs)	723	723 (100)
	2003–2007 (5 yrs)	435	435 (100)
	2002–2007 (6 yrs)	317	317 (100)
	2001–2007 (7 yrs)	244	244 (100)
	2000–2007 (8 yrs)	198	198 (100)
	1999–2007 (9 yrs)	115	115 (100)
	1998–2007 (10 yrs)	13	13 (100)

	Time Frame of EMR Data	FP Subjects, No.	Missing Diabetes Diagnosis Codes, No. (%)	Missing Use History of DM-Related Medications or Supplies, No. (%)	Missing Prior Abnormal Laboratory Reports, No. (%)
Control	2007 (1 yr)	4690	2263 (48)	24 (0.5)	2403 (51)
	2006–2007 (2 yrs)	4752	2039 (43)	21 (0.4)	2692 (57)
	2005–2007 (3 yrs)	4172	1767 (42)	16 (0.4)	2389 (57)
	2004–2007 (4 yrs)	3716	1625 (44)	9 (0.2)	2082 (56)
	2003–2007 (5 yrs)	3175	1408 (44)	6 (0.2)	1761 (55)
	2002–2007 (6 yrs)	2638	1168 (44)	5 (0.2)	1465 (56)
	2001–2007 (7 yrs)	1474	317 (22)	4 (0.3)	1153 (78)
	2000–2007 (8 yrs)	933	115 (12)	1 (0.1)	817 (88)
	1999–2007 (9 yrs)	569	38 (7)	0 (0)	531 (93)
	1998–2007 (10 yrs)	283	16 (6)	0 (0)	267 (94)

Abbreviations: DM, diabetes mellitus; EMR, electronic medical record; FP, false-positive; T1DM, type 1 diabetes mellitus.

Table 3

The Numbers of and Reasons that Contribute to False-Negative Subjects.

	Time Frame of EMR Data	FN Subjects, No.	Missing Diagnosis Codes for T2DM, No. (%)	Missing Prior Abnormal Laboratory Reports, No. (%)	Missing Use History of DM-Related Meds or Supplies, No. (%)
Case	2007 (1 yr)	681	384 (56)	137 (20)	160 (23)
	2006–2007 (2 yrs)	404	197 (49)	128 (32)	79 (20)
	2005–2007 (3 yrs)	304	133 (44)	122 (40)	49 (16)
	2004–2007 (4 yrs)	220	89 (40)	103 (47)	28 (13)
	2003–2007 (5 yrs)	154	56 (36)	80 (52)	18 (12)
	2002–2007 (6 yrs)	120	42 (35)	64 (53)	14 (12)
	2001–2007 (7 yrs)	78	27 (35)	44 (56)	7 (9)
	2000–2007 (8 yrs)	49	16 (33)	29 (59)	4 (8)
	1999–2007 (9 yrs)	27	9 (33)	18 (67)	0 (0)
	1998–2007 (10 yrs)	15	5 (33)	10 (67)	0 (0)
	Time Frame of EMR Data	**FN Subjects, No.**	**Absence of Enough Face-to-Face Visits, No. (%)**	**Missing Prior Laboratory Reports, No. (%)**	
Control	2007 (1 yr)	14119	2866 (20)	11253 (80)	
	2006–2007 (2 yrs)	10057	629 (6)	9428 (94)	
	2005–2007 (3 yrs)	7653	343 (4)	7310 (96)	
	2004–2007 (4 yrs)	5921	193 (3)	5728 (97)	
	2003–2007 (5 yrs)	4421	120 (3)	4301 (97)	
	2002–2007 (6 yrs)	3317	86 (3)	3231 (97)	
	2001–2007 (7 yrs)	2399	60 (3)	2339 (98)	
	2000–2007 (8 yrs)	1641	35 (2)	1606 (98)	
	1999–2007 (9 yrs)	1037	19 (2)	1018 (98)	
	1998–2007 (10 yrs)	481	7 (1)	474 (99)	

Abbreviations: DM, diabetes mellitus; EMR, electronic medical record; FN, false-negative; T2DM, type 2 diabetes mellitus.