Clinical Medicine & Research
Volume 5, Number 1: 1-7
©2007 Marshfield Clinic
http://www.clinmedres.org

Use of an Electronic Medical Record for the Identification of Research Subjects with Diabetes Mellitus

Russell A. Wilke, MD, PhD; Richard L. Berg, MS; Peggy Peissig, MBA; Terrie Kitchner;
Bozana Sijercic, MD; Catherine A. McCarty, PhD; and Daniel J. McCarty, PhD

Diabetes mellitus is a rapidly increasing and costly public health problem. Large studies are needed to understand the complex gene-environment interactions that lead to diabetes and its complications. The Marshfield Clinic Personalized Medicine Research Project (PMRP) represents one of the largest population-based DNA biobanks in the United States. As part of an effort to begin phenotyping common diseases within the PMRP, we now report on the construction of a diabetes case-finding algorithm using electronic medical record data from adult subjects aged ≥50 years living in one of the target PMRP ZIP codes. Based upon diabetic diagnostic codes alone, we observed a false positive case rate ranging from 3.0% (in subjects with the highest glycosylated hemoglobin values) to 44.4% (in subjects with the lowest glycosylated hemoglobin values). We therefore developed an improved case finding algorithm that utilizes diabetic diagnostic codes in combination with clinical laboratory data and medication history. This algorithm yielded an estimated prevalence of 24.2% for diabetes mellitus in adult subjects aged ≥50 years.

Keywords: Metformin; Natural language processing; Pharmacogenetics; Sulfonylurea

The current obesity epidemic represents a major international health problem.[1] Genetic markers may be the most efficient way to identify individuals at risk for obesity-related medical complications. One of the most costly obesity-related co-morbidities is diabetes mellitus (DM).[2] Hyperglycemia is the clinical hallmark of DM, but the etiology of this heterogeneous disorder likely involves multiple genetic and environmental interactions that ultimately result in alterations in insulin secretion, insulin action or both.[3,4] Large population-based cohorts will be needed to characterize the genetics of complex diseases such as DM.[5,6]

The Marshfield Clinic Personalized Medicine Research Project (PMRP) is a population-based DNA biobank developed to facilitate research in pharmacogenetics, genetic epidemiology and population genetics (www.mfldclin.edu/pmrp).[7] In 2003, the PMRP was mentioned in an article by Dr. Francis Collins and colleagues from the National Human Genome Research Institute as it relates to their identified grand challenge to "develop robust strategies for identifying the genetic contributions to disease and drug response."[8] Therefore, a PMRP Working Group was formed to select diseases for which electronic algorithms could be developed to classify exposure and outcome status using the electronic medical records contained within the database. The diseases represent a range of anticipated difficulty in using purely electronic methods to identify disease onset, disease progression and outcome. The first three diseases were selected from a list of diseases that are routinely screened for during routine health maintenance examinations in adults. Listed in order from expected greatest difficulty to least difficulty for electronic algorithms, the three diseases are: (1) glaucoma, (2) osteoporosis, and (3) DM. The purpose of the current study was to pilot the process of electronically and manually abstracting information from the electronic medical record of adults served by Marshfield Clinic to define DM specifically, so that the PMRP database could eventually be utilized for studies designed to characterize the genetic epidemiology and pharmacogenetics of this disease.

Methods

The current study protocol was approved by the Marshfield Clinic Institutional Review Board. The setting was a large multi-specialty group practice located in central Wisconsin.

Reprint Requests: Russell A. Wilke, MD, PhD, Center for Human Genetics, Marshfield Clinic Research Foundation, 1000 North Oak Avenue, Marshfield, WI 54449, Tel: 715-389-3885, Fax: 715-389-3808, E-mail: wilke.russell@mcrf.mfldclin.edu

Received: October 10, 2006
Revised: December 22, 2006
Accepted: January 8, 2007

DOI: 10.3121/cmr.2007.726

Grant Support: This project was funded by a grant from Marshfield Clinic Research Foundation.

The target population included residents within a single ZIP code (54449), encompassing the city of Marshfield (population 19,000). This ZIP code was selected because nearly everyone in the population seeks their health care through Marshfield Clinic, a fully integrated health care system with a long-standing comprehensive electronic medical record.[9] The target ZIP code was also one of 19 ZIP codes selected to recruit subjects for the Marshfield Clinic PMRP.[7]

Briefly, PMRP is a large biobank containing DNA and sera from approximately 19,000 Marshfield Clinic patients. Each PMRP participant has also provided informed consent allowing their genetic and serologic data to be linked to all available clinical data within their electronic medical record using a confidential and secure encryption process. PMRP therefore provides a unique opportunity to conduct very large genetic studies on a variety of common diseases.

Medical Record

Electronic medical records have been utilized at Marshfield Clinic since the 1960s, and the vast majority of patient records within this system have been electronic for over a decade. A variety of data are captured. One of the key features of the Marshfield Clinic electronic medical record is a Windows application called the combined medical record (CMR). CMR integrates data from all Marshfield Clinic facilities and cooperating hospitals, including Saint Joseph's Hospital (Marshfield). CMR includes indices to all events and encounters that patients have experienced within the Marshfield Clinic system of care, and it can be used to access all textual documentation such as office notes, operative reports and discharge summaries. CMR also includes comprehensive lists of patient problems, a summary of each clinic encounter (diagnoses and procedures), a variety of medication alerts, and online access to over a decade of laboratory and radiology results. Since nearly everyone residing in the target ZIP code for the current study receives their health care through Marshfield Clinic, this record is considered comprehensive.

Study Population

Subjects were considered eligible for this study based on the following criteria: (1) age 50 years or older, (2) alive on December 31, 2002, (3) seen at Marshfield Clinic between January 1, 2000 and December 31, 2002, and (4) residing in ZIP code 54449 (Marshfield). Electronic medical record data for the eligible subjects were searched to determine the presence (or absence) of diabetes diagnostic codes from the *International Classification of Diseases, Ninth Revision* (ICD-9 codes). These codes included primary diagnostic codes for diabetes (ICD-9 codes 250.00-250.92), and secondary diagnostic codes for diabetic neuropathy (ICD-9 code 357.2), retinopathy (ICD-9 codes 362.01-362.02) and nephropathy (ICD-9 code 583.81). For each potential study subject, clinical laboratory data were scanned electronically to identify relevant test results. These included all available glucose and glycosylated hemoglobin (HbA1c) values. Each glucose value was assumed to be random (i.e., non-fasting) unless otherwise specified. Maximum values were determined for each subject.

Medication History

We have previously utilized natural language processing (NLP) software to reconstruct complete retrospective medication use histories for all research subjects participating in the PMRP Biobank.[10] We have also shown previously that these data are amenable to electronic abstraction, and that they can be managed programmatically to yield high quality drug exposure histories in the context of lipid lowering therapy (e.g., 100% sensitive and 96% specific, with a precision of 95%).[11] In the current study, clinic records from

Table 1. Electronically abstracted text mention of glucose lowering medication* for the entire study cohort (n=8101).

	Diagnostic Code Available: Laboratory Data Available:	Yes Yes	Yes No	No Yes	No No	
Drug*						Total
I M S		532	1	33	10	576
I M —		125	0	39	5	169
I — S		128	1	25	3	157
I — —		167	3	334	25	529
— M S		47	0	6	2	55
— M —		45	0	38	2	85
— — S		29	0	28	2	59
— — —		335	9	5094	1033	6471
Total		1408	14	5597	1082	8101

* Drug code: I, insulin; M, metformin; S, sulfonylurea.

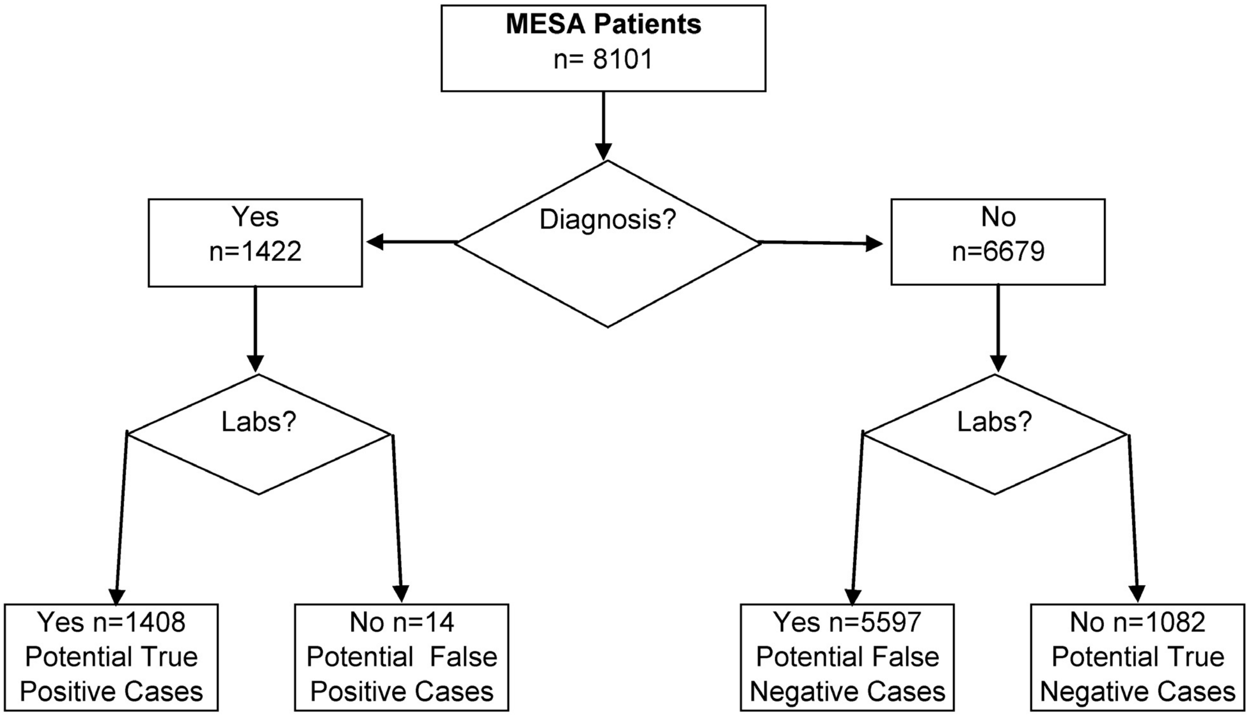

Figure 1. Initial electronic classification of study subjects based upon two criteria: first, the presence or absence of diabetic diagnostic codes, and second, the presence or absence of relevant clinical laboratory data (e.g., glucose levels and glycosylated hemoglobin [HbA1c] levels). This strategy produced four data bins. These bins contain 1408, 14, 5597 and 1082 study subjects, respectively (total study population 8101). MESA = Marshfield Epidemiologic Study Area.

all eligible subjects were re-interrogated electronically for text mention of three classes of glucose lowering medications. This involved the application of NLP software entitled FreePharma (Language & Computing; http://www.landc.be). All 8101 subject records were searched electronically to identify and catalogue dates for all text notes mentioning any sulfonylurea agent known to be commercially available within the past decade. This included four "first-generation" sulfonylureas (acetohexamide, chlorpropamide, tolbutamide, tolazamide) and three "second-generation" sulfonylureas (glimepiride, glipizide, glyburide). A similar approach was taken to identify all text notes containing mention of any therapeutic agent mapping to the generic drug names metformin (the only clinically approved glucose-lowering biguanide) and insulin (table 1).

Diagnostic Confirmation

A five-page data abstraction form was developed for use by trained research coordinators to manually abstract data related to DM diagnosis and treatment from the medical records. This form was used to collect demographic data and specific diabetes-related clinical information. For quality assurance, 10% of all manually abstracted records were re-abstracted by a second research coordinator, and discrepancies resolved by a licensed practicing physician. Research coordinators were asked to manually abstract data for three sets of subjects *with* electronically recognized diabetic diagnostic codes: (1) 100 subjects with the highest HbA1c, (2) 100 subjects with the lowest HbA1c, and (3) 14 specific subjects with records containing diabetic diagnostic codes but no corresponding laboratory data. Research coordinators were also asked to manually abstract data from a specific sample of subjects *without* electronic diabetic diagnostic codes: 72 subjects who had the most extreme glucose or HbA1c results. American Diabetes Association (ADA) diagnostic criteria were used to confirm the presence or absence of DM (i.e., fasting glucose ≥126 on two occasions or a single random glucose >200).

Results

The study population included 8101 patients who met the inclusion criteria. This number is comparable to the year 2000 US Census estimate (n = 7905) for this age group and ZIP code. All medical records from these study subjects were interrogated electronically for the presence of diagnostic codes associated with DM. Of the 8101 study subjects, 1422 (17.6%) subjects were found to have at least one diabetic diagnostic entry, i.e., either diabetes or a diabetic end organ complication (figure 1, left). The remaining medical records (n = 6679 study subjects) had no diabetic diagnostic entries (figure 1, right). Each of these two initial subsets (1422 with codes and 6679 without) is discussed separately below in the context of phenotyping accuracy.

Diagnostic Codes Present

Among the 1422 study subjects with at least one diabetic diagnostic code, 99% (1408 study subjects) had sufficient clinical laboratory data to either support or refute the

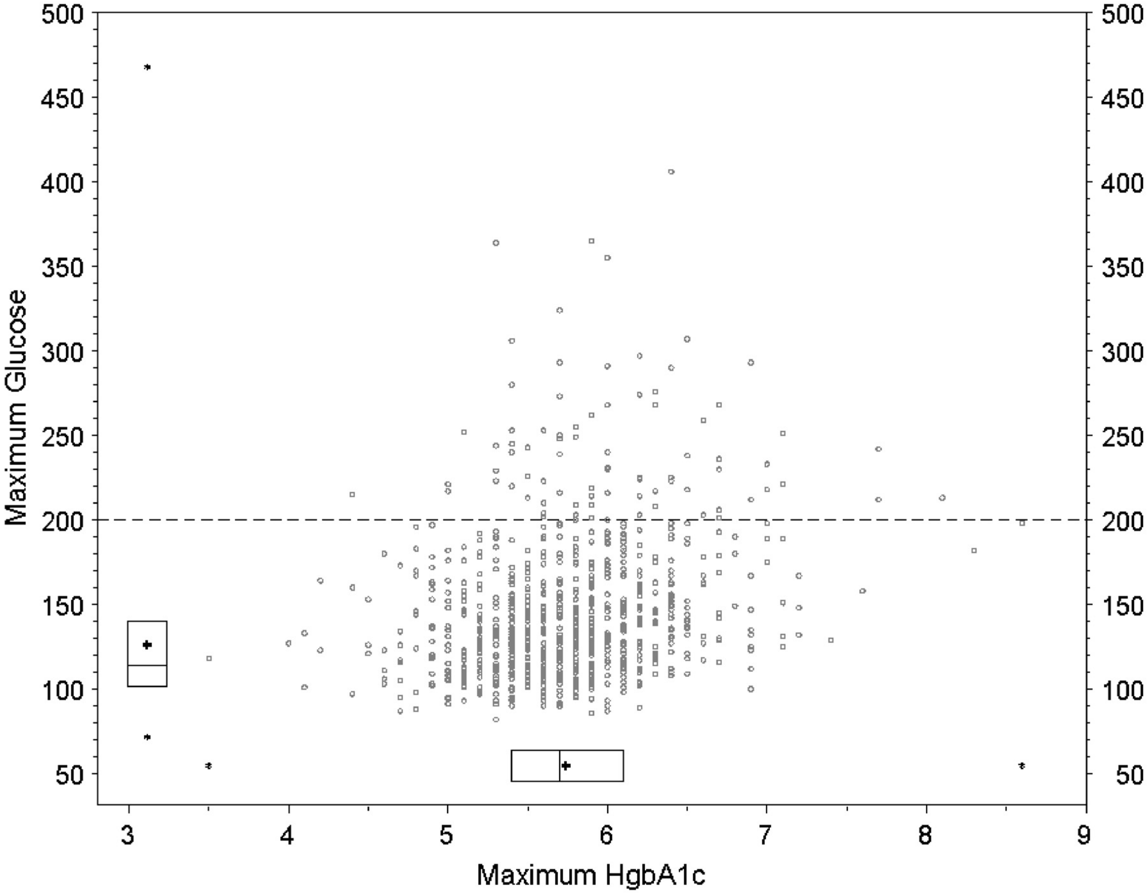

Figure 2. Graphic representation of the clinical laboratory data from bin 3 (n = 5597), as classified in the text (*potential* false negative cases) and illustrated in figure 1. Glucose levels *and* glycosylated hemoglobin (HbA1c) levels were abstracted electronically for all 5597 subjects in bin 3. The 854 patients who had at least one glucose level and at least one HbA1c level are shown as circles in the scatter plot. The box plots in the margins reflect all available data. The mean is shown as a "+" within boxes representing the 25th, 50th and 75th percentiles. Asterisks denote minimum and maximum. The dashed horizontal line indicates a glucose level 200 mg/dl.

diagnosis. These data include fasting glucose, random glucose or HbA1c levels. The1408 subject records containing these data represent potential true positive cases of DM (figure 1). Based upon manual data abstraction, we observed that diagnostic codes yielded a true positive rate for DM ranging from 97.0% (in 100 subjects with the highest HbA1c values) to 55.6% (in 100 subjects with the lowest HbA1c values). It should also be noted that our initial electronic screening strategy (e.g., diagnostic codes and laboratory data), as shown in figure 1, also yielded 14 potential false positive cases of DM (i.e., diabetic diagnostic codes without any supporting electronic laboratory data). Among these 14 potential false positive cases of DM, four were manually confirmed to be diabetic based upon treatment history or laboratory data not available electronically.

Diagnostic Codes Absent

Electronic interrogation of the entire medical record for each of the 8101 unique subjects in this study revealed that 6679 of these subjects had no diabetic diagnostic codes contained within their electronic medical record (figure 1, right side). Of these, 5597 (84%) had clinical laboratory data containing at least one glucose value or at least one HbA1c level. Since it was likely that some of these 5597 potential false negative cases were actually either undiagnosed diabetics or diabetics treated without a corresponding provider-entered diagnostic code, relevant clinical laboratory data were re-abstracted electronically for all 5597 subjects. These clinical laboratory data are summarized in figure 2. For both axes (glucose and HbA1c), the mean is represented by a "+" located within box plots corresponding to the 25th, 50th and 75th percentiles, respectively, for the entire dataset (n = 5597). The horizontal dashed line delineates a glucose level ≥200 mg/dl.

Within figure 2, only those 854 subjects found to have *both* a glucose level and an HbA1c level have been represented as circles in the scatter plot. Data were manually abstracted for 72 study subjects with the most extreme glucose and HbA1c values. Of these, 41 records contained a glucose value >200 mg/dl. All 41 records (100%) were manually confirmed as cases of DM.

Of the 8101 unique subjects in this study with no diabetic diagnostic codes (figure 1, right side), 1082 (16%) had no clinical laboratory data that could be used to discriminate between diabetic and non-diabetic (i.e., no glucose levels and no HbA1c levels). These 1082 subjects are assumed to be true negative cases (i.e., not diabetic). The design of this study (retrospective chart review) does not allow the discrimination of false negative cases within this specific sub-sample because the research subjects were neither interviewed nor examined during the conduct of the study. However, this population is known to be highly compliant with primary prevention screening visits.[12] Among the 5597 potential false negative case subjects with laboratory data but no diagnostic codes, 4477 (80%) were found to have at least one glucose level within 2 years. Based upon these observations, and the additional observation that patients residing in the target study ZIP code receive nearly all their healthcare (90% of outpatient visits, 95% of inpatient visits) through Marshfield Clinic,[9] it is reasonable to assume that the frequency of false negative cases would be low in the sub-sample of 1082 subjects with no relevant clinical laboratory data.

Prevalence Estimate

We propose the electronic case-finding algorithm shown in figure 3. The observations outlined above (*Diagnostic Codes Present versus Diagnostic Codes Absent*) suggest that the first branch point in this algorithm can be based upon diagnostic codes. The two subsequent branches of the algorithm then apply differential logic, reflecting the following two assumptions. First, in the situation where diabetic diagnostic codes are present, any purely electronic algorithm simply needs to *confirm* the diagnosis. This can be done by documenting either abnormal laboratory data (HbA1c>ULN, or glucose criteria established by the ADA) or treatment with one of three known medications used as first line therapy for DM. Conversely, in the situation where diabetic diagnostic codes are absent, the algorithm needs to *establish* the diagnosis. Since this latter step is more than simply confirmatory, the rightward arm of the algorithm needs to be sufficiently stringent to minimize (and, if possible, avoid altogether) false positive case assignment. Based upon the distribution of laboratory data observed in figure 2 (sub-sample with n=5597), we recommend that the identification of false positive case subjects within this sub-sample be made by first using the presence of an HbA1c test to suggest a reasonable clinical index of suspicion for DM, and then, second by accepting a maximum glucose value >200 mg/dl as diagnostic.

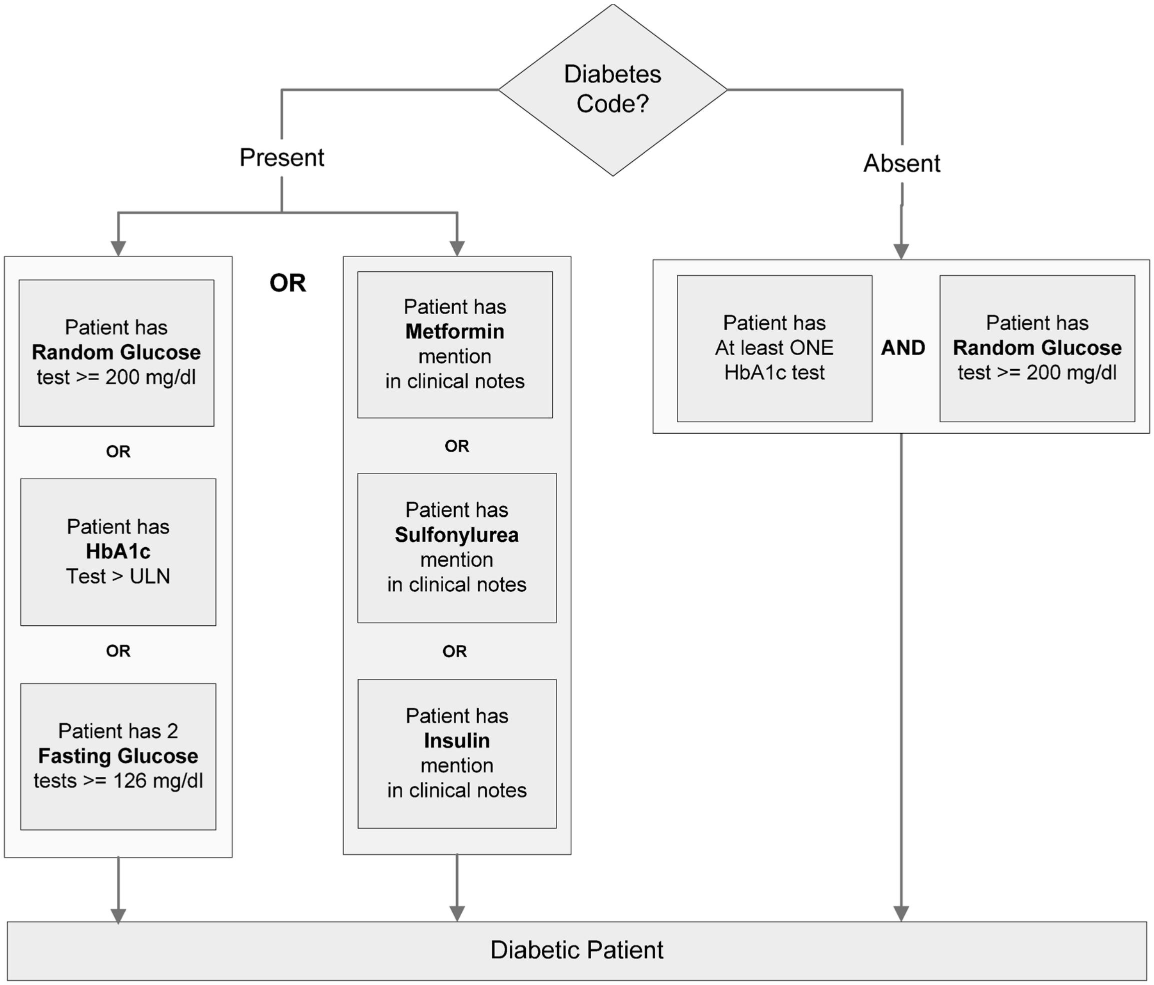
Diabetes Code?
Present
Absent
OR
Patient has Random Glucose test >= 200 mg/dl
OR
Patient has HbA1c Test > ULN
OR
Patient has 2 Fasting Glucose tests >= 126 mg/dl
Patient has Metformin mention in clinical notes
OR
Patient has Sulfonylurea mention in clinical notes
OR
Patient has Insulin mention in clinical notes
Patient has At least ONE HbA1c test
AND
Patient has Random Glucose test >= 200 mg/dl
Diabetic Patient

Figure 3. Proposed algorithm for identification of case subjects with diabetes mellitus (DM) in the Personalized Medicine Research Project (PMRP) database.

The final electronic algorithm was used to identify unique patients with DM. This electronic algorithm was applied to the entire study cohort (8101 adults aged ≥50 years and living in the target ZIP code), identifying 1960 (24.2%) unique subjects with DM.

Discussion

The current study outlines the construction of a case finding algorithm for DM. This algorithm utilizes diagnostic codes, clinical laboratory data and medication history to identify subjects with DM in a large patient database. Using diagnostic codes alone, we observed a high rate of false positive cases. Further confirmation is therefore required through clinical laboratory data (using current ADA diagnostic criteria[3]) or medication data (obtained by NLP[10]). By considering these additional data, the final algorithm reduces the frequency of false positive cases.

The final algorithm also reduces the frequency of false negative cases by identifying subjects with DM in the absence of a diabetic diagnostic code. However, this portion of the algorithm is conservative in that it requires the presence of an elevated random glucose level (≥200 mg/dl) specifically within the context of a subject record also containing at least one HbA1c value. We opted not to accept an elevated glucose level alone, since in the absence of diagnostic codes for diabetes, a random glucose value can be elevated for a variety of non-diagnostic reasons (e.g., steroid therapy or intravenous fluid replacement containing dextrose). Since the presence of at least one HbA1c test (whether normal or elevated) indicates an increased clinical index of suspicion for DM, an elevated random glucose level can be considered diagnostic in this context. Although stringent, our inclusion of a strategy to reduce false negative cases was necessary in this study population because the Centers for Disease Control and Prevention have estimated that a significant proportion of all adult diabetic subjects in the United States remain undiagnosed.[13]

Application of the final algorithm yielded an estimated DM prevalence of 24.2% for adults aged ≥50 years residing in the target ZIP code (i.e., the algorithm identified 1960 of the 8101 study subjects as diabetic case subjects). The prevalence of DM is highly associated with age, and our observation is consistent with previously published estimates.[13-15] This work adds to a growing body of literature supporting the utility of electronic medical records for case-finding specifically within the context of DM.[16-18] Further, the current study extends these observations through the development of an electronic algorithm that considers clinical laboratory data and medication history in addition to diagnostic codes. Since the target ZIP code characterized in the current study is located within the geographic region represented by the Marshfield Clinic PMRP database, the resulting algorithm will be useful for identifying DM cases in this database.

Outlook

The current study presents an electronic case-finding algorithm that can be used for the identification of research subjects with DM in the PMRP DNA biobank. It is important to note that DM is a clinically heterogeneous disorder, and that the current study does not discriminate between major forms of the disease. No effort was made to sub-classify subjects identified by this algorithm according to major type (e.g., type 1 versus type 2 diabetes), or minor type (e.g., maturity onset diabetes of the young or latent autoimmune diabetes in adults). It is anticipated that further phenotypic discrimination will be accomplished, on a study-by-study basis, during future applications of this algorithm using context-specific parameters defined by each study.

Acknowledgments

The authors thank Ms. Laura Lobner for data collection and quality assurance, Ms. Carla Rottscheit for data management and Ms. Ekta Sirohi for the electronic abstraction of medications. We also thank Marshfield Clinic Research Foundation for its support through the assistance of Alice Stargardt and Linda Weis in the preparation of this manuscript.

The Working Group for the Personalized Medicine Research Project Phenotyping Engine includes Dr. Philip Giampietro, Dr. Robert Greenlee, Dr. Catherine McCarty, Dr. Daniel McCarty, Ms. Peggy Peissig and Dr. Russell Wilke.

References

1. Mokdad AH, Ford ES, Bowman BA, Dietz WH, Vinicor F, Bales VS, Marks JS. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. JAMA 2003; 289:76-79.
2. Fontaine KR, Bartlett SJ. Access and use of medical care among obese persons. Obes Res 2000; 8:403-406.
3. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2006; 29(suppl 1):S43-S48.
4. Wilke RA, Jochen AL, Maas DL and IM O'Shaughnessy: Hypoglycemia and Diabetes Mellitus. In: Kutty K, Sebastian JL, Mewis BA, Berg DD, Kochar, MS, eds. Kochar's concise textbook of medicine. 3rd ed. Baltimore, MD: Williams & Wilkins; 1998.
5. Kaiser J. Biobanks. Population databases boom, from Iceland to the U.S. Science 2002; 298:1158-1161.
6. Davis RL, Khoury MJ. The journey to personalized medicine. Personalized Med 2005; 2:1-4.
7. McCarty C, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Personalized Med 2005; 2:49-79.
8. Collins FS, Green ED, Guttmacher AE, Guyer MS; US National Human Genome Research Institute. A vision for the future of genomics research. Nature 2003; 422:835-847.
9. Greenlee RT. Measuring disease frequency in the Marshfield Epidemiologic Study Area (MESA). Clin Med Res 2003; 1:273-280.
10. Sirohi E, Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. Pac Symp Biocomput 2005; 10:308-318.

11. Peissig P, Sirohi E, Berg RL, Brown-Switzer C, Ghebranious N, McCarty CA, Wilke RA. Construction of atorvastatin dose-response relationships using data from a large population-based DNA biobank. Basic Clin Pharmacol Toxicol 2007; 100:286-288.
12. McCarty CA, Chyou PH, Greenlee R, McCarty DJ, Gunderson P, Reding D. Differences in preventive screening rates in Wisconsin farm and non-farm resident women. WMJ 2003; 102:22-26.
13. National Diabetes Fact Sheet. Centers for Disease Control and Prevention Web site. Available at: http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2005.pdf. Accessed September 21, 2005.
14. Flegal KM, Carroll MD, Ogden CL, Johnson CL. Prevalence and trends in obesity among US adults, 1999-2000. JAMA 2002; 288:1723-1727.
15. Harris MI, Flegal KM, Cowie CC, Eberhardt MS, Goldstein DE, Little RR, Wiedmeyer HM, Byrd-Holt DD. Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in U.S. adults. The Third National Health and Nutrition Examination Survey, 1988-1994. Diabetes Care 1998; 21:518-524.
16. Hassey A, Gerrett D, Wilson A. A survey of validity and utility of electronic patient records in a general practice. BMJ 2001; 322:1401-1405.
17. Szeto HC, Coleman RK, Gholami P, Hoffman BB, Goldstein MK. Accuracy of computerized outpatient diagnoses in a Veterans Affairs general medicine clinic. Am J Manag Care 2002; 8:37-43.
18. Newton KM, Wagner EH, Ramsey SD, McCulloch D, Evans R, Sandhu N, Davis C. The use of automated data to identify complications and comorbidities of diabetes: a validation study. J Clin Epidemiol 1999; 52:199-207.

Author Affiliations

Russell A. Wilke, MD, PhD
Center for Human Genetics
Marshfield Clinic Research Foundation and
Department of Internal Medicine
Marshfield Clinic
1000 North Oak Avenue
Marshfield, Wisconsin 54449

Richard L. Berg, MS
Biomedical Informatics Research Center
Marshfield Clinic Research Foundation
1000 North Oak Avenue
Marshfield, Wisconsin 54449

Peggy Peissig, MBA
Biomedical Informatics Research Center
Marshfield Clinic Research Foundation
1000 North Oak Avenue
Marshfield, Wisconsin 54449

Terrie Kitchner
Center for Human Genetics
Marshfield Clinic Research Foundation
1000 North Oak Avenue
Marshfield, Wisconsin 54449

Catherine A. McCarty, PhD
Center for Human Genetics
Marshfield Clinic Research Foundation
1000 North Oak Avenue
Marshfield, Wisconsin 54449

Bozana Sijercic, MD
Department of Internal Medicine
Marshfield Clinic
1000 North Oak Avenue
Marshfield, Wisconsin 54449

Daniel J. McCarty, PhD
Marshfield Epidemiology Research Center
Marshfield Clinic Research Foundation
1000 North Oak Avenue
Marshfield, Wisconsin 54449