Published in final edited form as:
Pharmacoepidemiol Drug Saf. 2014 February ; 23(2): 111–118. doi:10.1002/pds.3557.

Distinguishing incident and prevalent diabetes in an electronic medical records database

Ronac Mamtani[1], **Kevin Haynes**[2], **Brian S Finkelman**[2], **Frank I Scott**[2], and **James D Lewis**[2]

[1]Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, United States

[2]Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA, United States

Abstract

Purpose—To develop a method to identify incident diabetes mellitus (DM) using an Electronic Medical Records (EMR) database, and test this classification by comparing incident and prevalent DM with common outcomes related to DM duration.

Methods—Incidence rates (IRs) of DM (defined as a first diagnosis or prescription) were measured in 3-month intervals through 36 months after registration in The Health Improvement Network, a primary care database, from 1994 to 2012. We used Joinpoint regression to identify the point where a statistically significant change in the trend of IRs occurred. Further analyses used this point to distinguish those likely to have incident (n=50,315) versus prevalent (n=28,337) DM. Incident and prevalent cohorts were compared using Cox regression for all-cause mortality, cardiovascular disease, diabetic retinopathy, diabetic nephropathy, and diabetic neuropathy. Analyses were adjusted for age, sex, smoking, obesity, hyperlipidemia, hypertension, and calendar year.

Results—Trends in DM incidence rates plateaued 9 months after registration (p=0.04). All cause-mortality was increased (HR 1.62, 95% CI 1.53–1.70) among patients diagnosed with DM prior to 9 months following registration (prevalent DM) compared to those diagnosed after 9 months (incident DM). Similarly, the risk of DM-related complications was higher in prevalent vs. incident DM patients [cardiovascular disease, HR 2.24 (2.08–2.40); diabetic retinopathy, HR 1.31 (1.24–1.38); diabetic nephropathy, HR 2.30 (1.95–2.72); diabetic neuropathy, HR 1.28 (1.16–1.41)].

Conclusion—Joinpoint regression can be used to identify patients with newly diagnosed diabetes within EMR data. Failure to exclude patients with prevalent DM can lead to exaggerated associations of DM-related outcomes.

Keywords

diabetes; incidence; bias; cohort studies; electronic medical records

INTRODUCTION

Electronic medical record (EMR) databases are widely used to study the epidemiology and outcomes of diabetes mellitus (DM).[1–3] Indeed, antidiabetic drugs are among the most widely studied classes in pharmacoepidemiologic research in the last decade.[4–6] When using

Corresponding author: Ronac Mamtani, MD MSCE, 16 Penn Tower, 3400 Spruce Street, Philadelphia, PA, 19104, ronac.mamtani@uphs.upenn.edu, phone: 215-615-1607, fax: 215-662-2432.

Conflicts of interest: no disclosures.

EMR to conduct pharmacoepidemiologic studies of DM, it is important to distinguish incident from prevalent DM. The identification of incident DM allows for an accurate assessment of important clinical outcomes related to DM duration. Furthermore, limiting studies to only newly diagnosed DM avoids bias from missing data on prior DM treatment (i.e., left censoring), which is particularly important when comparing the safety and effectiveness of alternative antidiabetic treatments. Thus, the ability to identify an incident DM cohort provides a defined period of follow-up after a first DM diagnosis, producing estimates of DM-associated outcomes and treatment effects that are free from bias due to left censoring.

Previous methods to identify incident DM in automated databases have relied on chart reviews[7,8] or more commonly case definitions requiring a minimum diabetes-free baseline period to minimize misclassification of prevalent cases as incident.[9–13] Chart reviews are often expensive and time-consuming. For the latter approach, several arbitrary baseline periods have been proposed, ranging from one month[13] to beyond 5 years.[9] Importantly, only two of the studies conducted statistical analyses to derive these periods (6 months and 5 years)[9,10], while none of the studies used these time points to compare resultant incident and prevalent DM with respect to complications of DM.

We therefore aimed to distinguish incident from prevalent DM using changes in trends in diabetes incidence rates, and test this classification by examining the association between incident and prevalent DM with common outcomes associated with longer duration of DM including death, cardiovascular disease, diabetic retinopathy, diabetic nephropathy, and diabetic neuropathy.

METHODS

Data source

We used data from The Health Improvement Network (THIN), an electronic medical records database that is representative of the broader United Kingdom population.[14] Data available in THIN include demographic information, medical diagnoses, lifestyle characteristics, and other clinical measurements recorded by general practitioners (GPs) during clinical practice. Medical diagnoses within the database are recorded using Read codes, the standard primary care classification system in the UK.[15] THIN also records all new and repeat prescriptions written by the GPs as the electronic record is used to generate these prescriptions.

The accuracy and completeness of THIN data is well documented, and the database has been used for epidemiological studies of several chronic diseases, including diabetes.[16–24] The database currently contains the electronic medical records of over 9 million patients, allowing for precise estimates of incidence rates of even rare outcomes.

Study design and population

We conducted a retrospective cohort study examining incidence rates of DM in THIN from 1994 to 2012. The source population included all subjects with acceptable records (e.g., permanently registered; no out of sequence year of birth or registration date; transfer out date not missing or invalid; year of birth not missing or invalid; sex information not missing) who registered in a THIN participating practice. To ensure high quality data recording, we excluded patient who registered in the database prior to the date that the practice started using the Vision management software.[25] Likewise, we also excluded patients with a date of the first recorded diagnosis of diabetes in THIN prior to the date of registration. The study protocol was approved by the University of Pennsylvania's Institutional Review Board and the United Kingdom's Scientific Review Committee.

Primary outcome and follow-up

The primary outcome was a diagnosis of diabetes mellitus, defined by the patient's first Read code consistent with DM or prescription for an oral antidiabetic (OAD) or insulin occurring after the registration date and without a diagnosis of DM prior to the date of registration. Follow-up time started on the patient's registration date and ended with the earliest occurrence of either the primary outcome or one of the following censoring events: 1) the date that the patient transferred out of a THIN practice; 2) the date of death, identified from the medical records file or the administration file; or 3) the end of each 3-month interval (see Statistical analyses).

Statistical analyses

Primary analysis—Incidence rates (IRs) and 95% confidence intervals of DM were measured in 3-month intervals through 36 months after registration in THIN, according to the method proposed by Lewis et al.[10] To describe trends in incidence rates, we used the Joinpoint Regression Program. Joinpoint software is widely used for the analysis of trend data, and is freely available from the Surveillance Research Program of the US National Cancer Institute (http://surveillance.cancer.gov/joinpoint/). In these analyses, piecewise regression models are used to identify time points at which a statistically significant change in the trend of IRs occurred.[26] To arrive at the final model, the Joinpoint program uses a sequence of permutation tests to select the best-fitting point. P-values for this point are estimated using Monte Carlo methods, adjusted for multiple comparisons through the Bonferroni correction.[26]

Secondary analysis—Secondary analyses used this time point to distinguish patients likely having incident versus prevalent DM. Incident and prevalent cohorts were then compared using Cox regression for the following outcomes: all-cause mortality, cardiovascular disease (myocardial infarction, acute coronary syndrome, congestive heart failure, or stroke), diabetic retinopathy, diabetic nephropathy, diabetic neuropathy, and receipt of a first prescription for an oral antidiabetic drug (metformin, sulfonylurea, thiazolidinedione, acarbose, meglitinide, or incretin mimetic) or insulin. The presence of these outcomes was defined using the first Read code consistent with these diagnoses or drug code for diabetes medication prescriptions, as recorded by general practitioners. In these separate analyses, follow-up started on the date of the first DM diagnosis and ended with the earliest of the following: the secondary outcome of interest, transfer out of practice, death, or the last date for data collection by the practice.

Analyses were adjusted for potential confounders including age, sex, smoking (ever versus never), hyperlipidemia, hypertension, calendar year, and obesity (BMI ≥ 30). Covariates were selected based upon the established association of each of these variables with the risk of death and common diabetes outcomes. Potential confounders were measured using all available data prior to or within 7 days after the DM diagnosis date, with exception of smoking and use of diabetes medications.[27] Smoking status was measured using data recorded at any time before or during follow-up. Receipt of an oral antidiabetic or insulin prescription was assessed two or more weeks prior to a first DM diagnosis. Hypertension and hyperlipidemia were defined exclusively using medical codes for diagnosis, while obesity was defined using medical codes and recorded height and weight.

Because BMI values were missing in 14% and 28% of the incident and prevalent cohorts, respectively, linear regression was used to multiply impute missing BMI values. To account for the variability between imputations, standard errors were adjusted according to the method proposed by Rubin et al.[28] In a sensitivity analysis, we determined the potential bias from a missing or incompletely measured confounder on the observed association between

diabetes duration and diabetes-related complications by the method proposed by Lin et al.[29] In this analysis, we re-estimated the relative risk of cardiovascular disease, a common DM-related complication, in the prevalent versus incident DM cohort after adjustment for a hypothetical confounder.

A sensitivity analysis was performed to determine whether incidence rates of DM changed if only DM diagnostic codes without oral antidiabetic drug (OAD) or insulin codes were used to identify new diagnoses of diabetes. In this analysis, follow-up timed ended on the date of the patient's first diagnostic code for DM or censoring event as previously described. Joinpoint analysis was repeated using this alternative method of identifying patients with DM.

The proportional hazards assumptions were met for all models. All statistical tests were two-sided and tested at the 5% level of significance, using STATA version 12.1 (StataCorp, College Station, TX).

RESULTS

We included 3,700,388 patients who registered in THIN practices with Vision software (Figure 1). Mean follow-up time for the cohort was 2.5 years. Incident diagnoses of diabetes were highest during the first 3 months following registration, and subsequently declined toward a baseline (Figure 2). Joinpoint analyses demonstrated that the trends in the incidence rates of newly diagnosed DM plateaued 9 months (p=0.04) after registration (Figure 2). Subsequent analyses used this time point to categorize DM patients as incident and prevalent DM.

The prevalent cohort included 28,337 patients with a first DM diagnosis within 9 months of registration, while the incident cohort included 50,315 patients with a first DM diagnosis more than 9 months following registration (Figure 1). After a first DM diagnosis, the mean follow-up time in the cohort was 1.8 years, while the maximum follow-up time was 16 years. Compared to the prevalent cohort, DM patients in the incident cohort were younger (median age 51 vs. 53 years) and more likely to have obesity (38.1% vs. 28.7%), hypertension (29.9% vs. 22.5%), and hyperlipidemia (11.4% vs. 6.1%) (Table 1). In both cohorts, receipt of a prescription for an oral antidiabetic drug (OAD) or insulin two or more weeks prior to a first DM diagnosis was uncommon (2.1% and 0.5%, incident cohort; 4.6% and 1.6%, prevalent cohort). Of note, half of the prevalent cohort had their first DM diagnosis within 1 month of registration, such that there was extremely limited opportunity for a prior prescription for an OAD or insulin.

When we used bivariate analysis to assess the relative hazards of death between the cohorts, all cause-mortality was increased (HR = 1.95, 95% CI = 1.86–2.05) among patients diagnosed with DM prior to 9 months following registration (prevalent cohort) compared to those diagnosed after 9 months following registration (incident cohort) (Table 2). The fully adjusted (after adjustment for age, sex, smoking, hyperlipidemia, hypertension, obesity, and calendar year) hazard ratio (HR) for the prevalent relative to incident cohort was 1.62 (95% CI = 1.53–1.70).

In analyses that examined the relative hazards of incident DM complications (Table 2), we observed an increased risk of cardiovascular disease in DM patients classified as prevalent relative to incident (fully adjusted: HR = 2.24, 95% CI 2.08–2.40). Similarly, the risk of other DM complications including diabetic retinopathy, diabetic nephropathy, and diabetic neuropathy were higher in the prevalent compared to the incident cohort [fully adjusted: diabetic retinopathy, HR = 1.31 (1.24–1.38); diabetic nephropathy, HR = 2.30 (1.95–2.72); diabetic neuropathy, HR = 1.28 (1.16–1.41)] (Table 2).

Additionally, we compared time to a first prescription of an OAD or insulin among DM patients in the two cohorts. In these analyses, DM patients categorized as prevalent were more likely to initiate OAD or insulin therapy (fully adjusted: HR = 1.62, 95% CI = 1.59–1.67; than those categorized as incident.

In a sensitivity analysis using medical diagnoses without prescription data to identify new diagnoses of DM, we observed nearly identical incidence rates of DM (Figure 3). Consistent with our primary analysis, incidence rates of DM plateaued 9 months (p=0.02) after registration.

DISCUSSION

Among this large cohort of patients registering in the THIN database, we show that changes in trends in diabetes incidence can help to distinguish incident from prevalent DM. Specifically, incidence rates of a first diagnosis of DM reach a baseline approximately 9 months following a patient's registration. Furthermore, patients with a first diagnosis of DM within 9 months of registration with a GP have higher mortality rates and higher rates of complications associated with long term DM. These data suggest that exclusion of patients with a DM diagnosis prior to 9 months is necessary to more accurately identify newly diagnosed DM and to conduct unbiased studies of outcomes where duration of DM could be an important confounder.[30]

Although electronic medical record (EMR) databases can be used to identify new cases of DM for large populations, they can be subject to bias due to misclassification. Misclassification of prevalent cases as incident cases is more likely to occur if appropriate amounts of follow-time prior to a first diagnosis of diabetes are not excluded at baseline.[10] The impact of such bias depends on the outcomes under study. DM is a chronic disease and duration of disease is a major determinant of the risk of several DM-specific outcomes.[31–36] Thus, the ability to reliably identify a first diagnosis of DM also allows for a more accurate assessment of DM duration effects.

Available data to guide researchers in the identification of patients with new DM have been inconsistent, and limited mostly to administrative data sources.[9,11–13,37,38] Each of these studies has excluded different amounts of follow-up time prior to a first DM diagnosis in an attempt to remove the pool of patients with prevalent DM. For example, some studies using United States claims data have used a baseline period of as short as 1 month[13] or no baseline period,[37] while others using Canadian claims data employed time periods ranging from 2 years to more than 5 years.[9,11,12] Only two studies statistically derived the baseline period used for exclusion of patients with potentially prevalent DM.[9,10] Asghari et al used retrograde survival function to show that a 5-year diabetes-free clearance period is reliable to identify new DM.[9] Consistent with our data, Lewis et al used data from a related primary care database, the General Practice Research Database, to demonstrate that the incidence rates of several chronic conditions including diabetes generally approach a baseline by months 10–12.[10]

There are several important strengths of our study. We hypothesized that failure to exclude an appropriate DM-free baseline period may over represent prevalent cases leading to biased associations between DM and clinical outcomes related to DM duration. In our study, we reconfirmed the findings reported by Lewis et al,[10] and, by comparing prevalent relative to incident DM through the application of a baseline exclusion period, we demonstrated significant differences in the risk of death and other DM complications including cardiovascular disease, retinopathy, neuropathy, and nephropathy, lending support to our

hypothesis. Finally, the large size and sufficient follow-up of the cohorts in these analyses is reflected by the precision of the incidence rates and risk estimates observed.

Nonetheless, this study has several limitations. We did not attempt to validate our diabetes diagnostic codes through direct query of treating physicians or review of consultant letters as validation methods such as these often add substantial costs and time delay. However, the positive predictive value of diagnostic codes for diabetes in a related database (General Practice Research Database) for which there is overlap in practices with THIN and that uses the same electronic medical record software for data collection was recently reported to be 98%.[39] Furthermore, the incidence rate of diabetes in our cohort after 9 months following registration (months 9 – 36, IR = 295–370 per 100,000 person-years) was comparable with that previously reported in the UK (1994 – 2003, IR = 269 per 100,00 person years).[40]

We assumed that patients with a first DM diagnosis shortly after registration with a general practitioner (GP) had prevalent DM. Additionally, we excluded patients with a first DM diagnosis before registration given the high likelihood that a diagnosis from this period represents prevalent DM. However, since we used time after registration to identify new cases of disease, we expect some degree of misclassification between incident and prevalent DM. Clearly, not every patient with a diabetes diagnosis prior to 9 months of enrollment in THIN will have had prevalent DM. To test this hypothesis, we extracted data on preventive medicine visits among diabetes patients classified as prevalent using our algorithm. Among this cohort (N=28,337) we found that 15% (n=4,180) had medical codes (new patient screen, new registration consultation, diabetic register, and diabetic consultation) indicating a preventive medicine consultation and/or new patient screen at the time of or prior to their first diabetes diagnosis suggesting that these patients could have incident diabetes. In our study, such patients would be misclassified as having prevalent diabetes. However, use of these codes is likely incomplete and as such it is impossible to quantify the proportion of patients with diabetes in the first 9 months after registration who truly have prevalent diabetes. However, the main purpose of the proposed methods is to exclude prevalent cases from the incident cohort, not to assure capture of all incident patients. Therefore, to avoid such misclassification, researchers who wish to conduct a study comparing prevalent to incident diabetes could limit the prevalent cohort to subjects with a diabetes diagnosis prior to registration and limit the incident cohort to those who are diagnosed 9 or more months after registration.

In a sensitivity analysis using only diagnostic codes for diabetes without prescription data, we found that the change in trends in DM incidence occurred at the same time point (i.e., 9 months) as observed in our primary analysis. Of note, few patients categorized as prevalent had prior use of an oral antidiabetic drug (<5%) or insulin (<2%). Therefore, at least for the diagnosis of DM, it does not appear necessary to include prescription data when selecting a time period to exclude for the purpose of identifying incident diagnoses. It is therefore likely that our method can be applied in other datasets that lack prescription information. Whether the same would apply to other chronic diseases is currently unknown.

The differences in demographic features and cardiovascular risk factors between the cohorts provide further evidence of the clinical relevance of distinguishing incident from prevalent diabetes using this methodology. Data on variables known to be associated with both diabetes duration and diabetes-related clinical outcomes, including hypertension, hyperlipidemia, smoking status, and BMI, are available in THIN and have been adjusted for in this cohort study. We did not adjust for hemoglobin A1c (HbA1c) as baseline data were missing in nearly 60% of subjects. Furthermore, because many datasets lack HbA1c information, we chose not to focus on laboratory values. However, as in any observational study, our study is subject to unmeasured confounding from variables unavailable in the

database as well as residual confounding from variables that have been measured with insufficient detail. For example, the prevalence of hypertension (22–30%), obesity (28–38%), and hyperlipidemia (6–11%) in our cohorts is lower than expected in a UK diabetes population (HTN: 34–54%, OB: 52%, HL: 30–47%).[41,42] The low prevalence of these variables could be due to the exclusion of patients diagnosed with diabetes prior to registration. Such patients have a greater opportunity to develop and be diagnosed with hypertension, obesity, and hyperlipidemia. Additionally, the potential mixing of prevalent and incident diabetes among those diagnosed in the first 9 months after registration may artificially lower the prevalence of these variables. Even if there is under recording of these risk factors for heart disease, we found that a missing or incompletely measured confounder would need to be highly prevalent (~%80) and strongly associated (HR ~ 3) with a DM-related complication such as cardiovascular disease (CVD) to nullify the observed association between diabetes duration and CVD (Table 3).

Diabetes duration is associated with several clinical outcomes and is likely an important confounder in many pharmacoepidemiology studies. Identification of newly diagnosed diabetes facilitates measurement of diabetes duration. We have described a simple approach to distinguish incident and prevalent diabetes and have demonstrated the importance of this categorization by assessing its effect on the risk of these clinical outcomes. The method can be easily implemented in any healthcare database where the classification of incident and prevalent disease is important.

Acknowledgments

Funding: This research was supported by the National Institutes of Health (grant number K12 CA 076931 to RM, 1F30HL115992-01 to BSF, K08-DK095951-01 to FIS, UL1-RR024134 to KH and JDL, and K24-DK078228 to JDL).

References

1. Hippisley-Cox J, Pringle M. Prevalence, care, and outcomes for patients with diet-controlled diabetes in general practice: cross sectional survey. Lancet. Jul-Aug;2004 364(9432):423–428. [PubMed: 15288740]
2. Reed M, Huang J, Graetz I, et al. Outpatient electronic health records and the clinical care and outcomes of patients with diabetes mellitus. Annals of internal medicine. Oct 2; 2012 157(7):482–489. [PubMed: 23027319]
3. Nichols GA, Desai J, Elston Lafata J, et al. Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: the SUPREME-DM project. Preventing chronic disease. 2012; 9:E110. [PubMed: 22677160]
4. Bosco JL, Antonsen S, Sorensen HT, Pedersen L, Lash TL. Metformin and incident breast cancer among diabetic women: a population-based case-control study in Denmark. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive. Oncology. Jan; 2011 20(1):101–111.
5. Azoulay L, Yin H, Filion KB, et al. The use of pioglitazone and the risk of bladder cancer in people with type 2 diabetes: nested case-control study. BMJ. 2012; 344:e3645. [PubMed: 22653981]
6. Graham DJ, Ouellet-Hellstrom R, MaCurdy TE, et al. Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone. JAMA: the journal of the American Medical Association. Jul 28; 2010 304(4):411–418. [PubMed: 20584880]
7. Bobo WV, Cooper WO, Stein CM, et al. Positive predictive value of a case definition for diabetes mellitus using automated administrative health data in children and youth exposed to antipsychotic drugs or control medications: a Tennessee Medicaid study. BMC medical research methodology. 2012; 12:128. [PubMed: 22920280]

8. Ho ML, Lawrence N, van Walraven C, et al. The accuracy of using integrated electronic health care data to identify patients with undiagnosed diabetes mellitus. Journal of evaluation in clinical practice. Jun; 2012 18(3):606–611. [PubMed: 21332609]
9. Asghari S, Courteau J, Carpentier AC, Vanasse A. Optimal strategy to identify incidence of diagnostic of diabetes using administrative data. BMC medical research methodology. 2009; 9:62. [PubMed: 19715586]
10. Lewis JD, Bilker WB, Weinstein RB, Strom BL. The relationship between time since registration and measured incidence rates in the General Practice Research Database. Pharmacoepidemiology and drug safety. Jul; 2005 14(7):443–451. [PubMed: 15898131]
11. Blanchard JF, Ludwig S, Wajda A, et al. Incidence and prevalence of diabetes in Manitoba, 1986–1991. Diabetes care. Aug; 1996 19(8):807–811. [PubMed: 8842595]
12. Hux JE, Ivis F, Flintoft V, Bica A. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. Diabetes care. Mar; 2002 25(3):512–516. [PubMed: 11874939]
13. Kudyakov R, Bowen J, Ewen E, et al. Electronic health record use to classify patients with newly diagnosed versus preexisting type 2 diabetes: infrastructure for comparative effectiveness research and population health management. Population health management. Feb; 2012 15(1):3–11. [PubMed: 21877923]
14. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Informatics in primary care. 2011; 19(4):251–255.
15. Chisholm J. The Read clinical classification. BMJ. Apr 28.1990 300(6732):1092. [PubMed: 2344534]
16. Gonzalez EL, Johansson S, Wallander MA, Rodriguez LA. Trends in the prevalence and incidence of diabetes in the UK: 1996–2005. Journal of epidemiology and community health. Apr; 2009 63(4):332–336. [PubMed: 19240084]
17. Martin-Merino E, Fortuny J, Rivero E, Garcia-Rodriguez LA. Validation of diabetic retinopathy and maculopathy diagnoses recorded in a U.K. primary care database. Diabetes care. Apr; 2012 35(4):762–767. [PubMed: 22357184]
18. Hall GC, McMahon AD, Carroll D, Home PD. Macrovascular and microvascular outcomes after beginning of insulin versus additional oral glucose-lowering therapy in people with type 2 diabetes: an observational study. Pharmacoepidemiology and drug safety. Mar; 2012 21(3):305–313. [PubMed: 22271442]
19. Mulnier HE, Seaman HE, Raleigh VS, Soedamah-Muthu SS, Colhoun HM, Lawrenson RA. Mortality in people with type 2 diabetes in the UK. Diabetic medicine: a journal of the British Diabetic Association. May; 2006 23(5):516–521. [PubMed: 16681560]
20. Denburg MR, Haynes K, Shults J, Lewis JD, Leonard MB. Validation of The Health Improvement Network (THIN) database for epidemiologic studies of chronic kidney disease. Pharmacoepidemiology and drug safety. Nov; 2011 20(11):1138–1149. [PubMed: 22020900]
21. Lo Re V 3rd, Haynes K, Forde KA, Localio AR, Schinnar R, Lewis JD. Validity of The Health Improvement Network (THIN) for epidemiologic studies of hepatitis C virus infection. Pharmacoepidemiology and drug safety. Sep; 2009 18(9):807–814. [PubMed: 19551699]
22. Lewis JD, Brensinger C, Bilker WB, Strom BL. Validity and completeness of the General Practice Research Database for studies of inflammatory bowel disease. Pharmacoepidemiology and drug safety. Apr-May;2002 11(3):211–218. [PubMed: 12051120]
23. Mamtani R, Haynes K, Bilker WB, et al. Association between longer therapy with thiazolidinediones and risk of bladder cancer: a cohort study. Journal of the National Cancer Institute. Sep 19; 2012 104(18):1411–1421. [PubMed: 22878886]
24. Haynes K, Forde KA, Schinnar R, Wong P, Strom BL, Lewis JD. Cancer incidence in The Health Improvement Network. Pharmacoepidemiology and drug safety. Aug; 2009 18(8):730–736. [PubMed: 19479713]
25. Larsson SC, Orsini N, Brismar K, Wolk A. Diabetes mellitus and risk of bladder cancer: a meta-analysis. Diabetologia. Dec; 2006 49(12):2819–2823. [PubMed: 17021919]

26. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Statistics in medicine. Feb 15; 2000 19(3):335–351. [PubMed: 10649300]
27. Brunelli SM, Gagne JJ, Huybrechts KF, et al. Estimation using all available covariate information versus a fixed look-back window for dichotomous covariates. Pharmacoepidemiology and drug safety. May; 2013 22(5):542–550. [PubMed: 23526818]
28. Rubin, D. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
29. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics. Sep; 1998 54(3):948–963. [PubMed: 9750244]
30. Suissa S, Azoulay L. Metformin and the risk of cancer: time-related biases in observational studies. Diabetes care. Dec; 2012 35(12):2665–2673. [PubMed: 23173135]
31. Fox CS, Sullivan L, D'Agostino RB Sr, Wilson PW. The significant effect of diabetes duration on coronary heart disease mortality: the Framingham Heart Study. Diabetes care. Mar; 2004 27(3): 704–708. [PubMed: 14988289]
32. Brun E, Nelson RG, Bennett PH, et al. Diabetes duration and cause-specific mortality in the Verona Diabetes Study. Diabetes care. Aug; 2000 23(8):1119–1123. [PubMed: 10937508]
33. Banerjee C, Moon YP, Paik MC, et al. Duration of diabetes and risk of ischemic stroke: the Northern Manhattan Study. Stroke; a journal of cerebral circulation. May; 2012 43(5):1212–1217.
34. Klein R, Klein BE, Moss SE, Davis MD, DeMets DL. The Wisconsin epidemiologic study of diabetic retinopathy. III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years. Archives of ophthalmology. Apr; 1984 102(4):527–532. [PubMed: 6367725]
35. Adler AI, Stevens RJ, Manley SE, Bilous RW, Cull CA, Holman RR. Development and progression of nephropathy in type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS 64). Kidney international. Jan; 2003 63(1):225–232. [PubMed: 12472787]
36. Young MJ, Boulton AJ, MacLeod AF, Williams DR, Sonksen PH. A multicentre study of the prevalence of diabetic peripheral neuropathy in the United Kingdom hospital clinic population. Diabetologia. Feb; 1993 36(2):150–154. [PubMed: 8458529]
37. Hebert PL, Geiss LS, Tierney EF, Engelgau MM, Yawn BP, McBean AM. Identifying persons with diabetes using Medicare claims data. American journal of medical quality: the official journal of the American College of Medical Quality. Nov-Dec;1999 14(6):270–277. [PubMed: 10624032]
38. Harris SB, Glazier RH, Tompkins JW, et al. Investigating concordance in diabetes diagnosis between primary care charts (electronic medical records) and health administrative data: a retrospective cohort study. BMC health services research. 2010; 10:347. [PubMed: 21182790]
39. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. The British journal of general practice: the journal of the Royal College of General Practitioners. Mar; 2010 60(572):e128–136. [PubMed: 20202356]
40. Hippisley-Cox, J. Diabetes in the United Kingdom: Analysis of QRESEARCH data. 2007. http://www.qresearch.org/Public_Documents/DataValidation/Diabetes%20in%20the%20UK%20analysis%20of%20QRESEARCH%20data.pdf
41. Daousi C, Casson IF, Gill GV, MacFarlane IA, Wilding JP, Pinkney JH. Prevalence of obesity in type 2 diabetes in secondary care: association with cardiovascular risk factors. Postgraduate medical journal. Apr; 2006 82(966):280–284. [PubMed: 16597817]
42. Samaranayaka S, Gulliford MC. Trends in cardiovascular risk factors among people with diabetes in a population based study, Health Survey for England 1994–2009. Primary care diabetes. May 16.2013

Key points

- Changes in trends in diabetes incidence rates can help identify incident diabetes in an electronic medical records database.
- Incidence rates of diabetes plateaued by 9 months following a patient's registration with their general practitioner.
- Patients with a first diagnosis of diabetes within 9 months of registration have higher mortality rates and higher rates of complications associated with longer duration of diabetes.
- Exclusion of follow-up time prior to 9 months is necessary to accurately identify patients with incident diabetes and to conduct unbiased studies of outcomes where duration of diabetes is an important confounder.

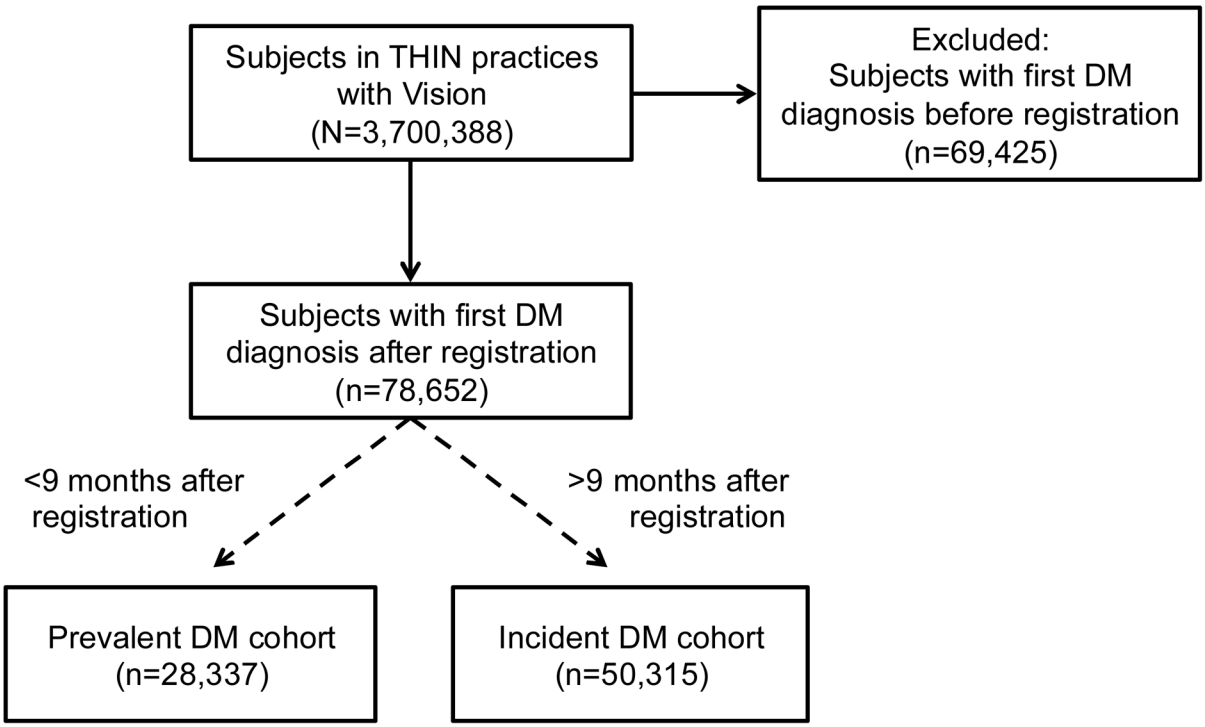

Figure 1.
Study flow diagram
A retrospective cohort study was conducted among subjects in THIN examining incidence rates of diabetes mellitus (DM) following registration in the database. Joinpoint regression was used to distinguish those likely to have incident (n=50,315) vs prevalent (n=28,337). Further analyses compared incident and prevalent DM cohorts with outcomes related to DM duration including all-cause mortality, cardiovascular disease, diabetic retinopathy, diabetic nephropathy, and diabetic neuropathy.

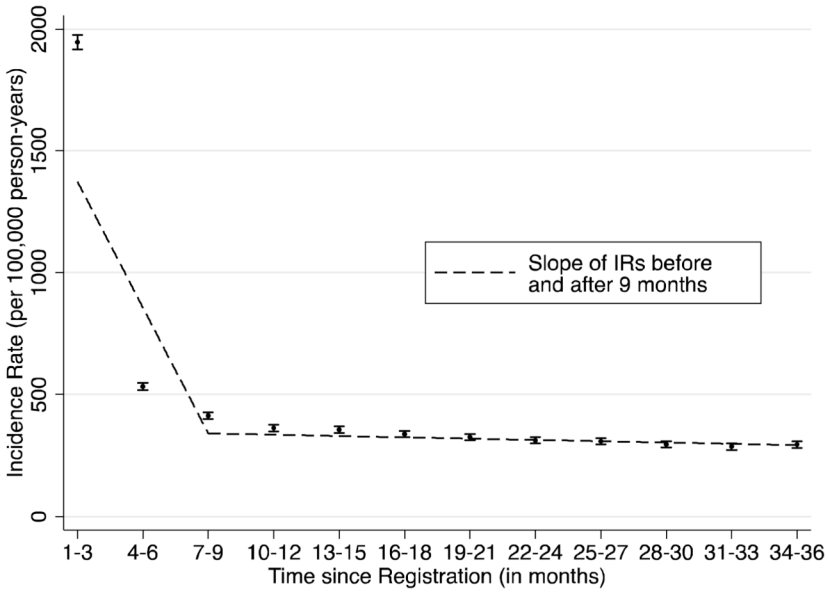

Figure 2.

Incidence of diabetes after registration in THIN

IR = incidence rate

IRs of diabetes were measured in 3-month intervals through 36 months after registration in The Health Improvement Network (THIN) database. Joinpoint regression was used to identify the point at which a change in the linear slope of IRs occurred. A diagnosis of diabetes mellitus was defined by the patient's first Read code consistent with DM or prescription for an oral antidiabetic (OAD) or insulin occurring after the registration date.

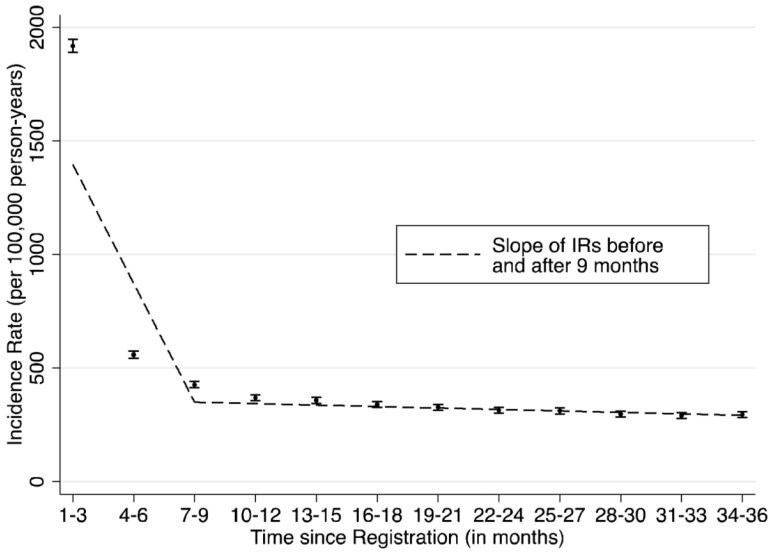

Figure 3.

Sensitivity analysis using diabetes diagnostic codes without medication prescriptions to measure incidence of diabetes after registration in THIN

IR = incidence rate

IRs of diabetes, using diabetes diagnostic codes without diabetes drug codes, were measured in 3-month intervals through 36 months after registration in The Health Improvement Network (THIN) database. Joinpoint regression was used to identify the point at which a change in the linear slope of IRs occurred.

Table 1

Demographics of the Incident and Prevalent Diabetes Cohorts [§]

	Incident DM[*] (n=50,315)	Prevalent DM[†] (n=28,337)
Age, y, median (IQR)	51 (37–66)	53 (36–68)
Sex, No. (% Male)	23,726 (47.2)	14,355 (50.7)
Smoking, No. (% Ever)	27,639 (54.9)	14,117 (49.8)
BMI, kg/m^2, No. (%)		
<30	24,027 (47.8)	12,331 (43.5)
≥30	19,159 (38.1)	8,146 (28.7)
Missing	7,129 (14.2)	7,869 (27.7)
Hypertension, No. (%)	15,030 (29.9)	6,383 (22.5)
Hyperlipidemia, No. (%)	5,738 (11.4)	1,789 (6.3)
Baseline OAD, No. (%) [#]	1,069 (2.1)	1,292 (4.6)
Baseline insulin, No. (%) [#]	243 (0.5)	454 (1.6)

DM = diabetes mellitus, IQR = interquartile range, BMI = body mass index, OAD = oral antidiabetic drug.

[§] All comparisons have p-values <0.01

[*] Patients diagnosed with DM after 9 months following registration

[†] Patients diagnosed with DM within 9 months following registration

[#] Receipt of at least one prescription for an oral antidiabetic drug (metformin, sulfonylurea, thiazolidinedione, acarbose, meglitinide, or incretin mimetic) or insulin two or more weeks prior to a first DM diagnosis

Table 2

Relative Hazard of Death, Diabetes-Related Complications, and DM Medication in the Prevalent vs Incident Diabetes Cohort

	Events	Person- years	Unadjusted (HR, 95% CI)	Fully adjusted* (HR, 95% CI)
Death	6636	87465025	1.95 (1.86–2.05)	1.62 (1.53–1.70)
DM Complications				
Cardiovascular disease	3525	72854261	2.10 (1.96–2.24)	2.24 (2.08–2.40)
Diabetic retinopathy	5656	80972693	1.37 (1.33–1.40)	1.31 (1.24–1.39)
Diabetic nephropathy	630	86327597	2.07 (1.77–2.42)	2.31 (1.95–2.73)
Diabetic neuropathy	1656	85052378	1.28 (1.16–1.41)	1.33 (1.19–1.47)
Rx for first oral antidiabetic or insulin†	31143	47192080	1.53 (1.50–1.57)	1.62 (1.59–1.67)

Rx = prescription, HR = hazard ratio, CI = confidence interval

*Adjusted for age, sex, smoking (ever vs never), calendar year, and histories of hyperlipidemia, hypertension, and obesity (BMI ≥ 30).

†Defined as time from the patient's first diabetes diagnosis to receipt of at least one prescription for an oral antidiabetic drug (metformin, sulfonylurea, thiazolidinedione, acarbose, meglitinide, or incretin mimetic) or insulin.

Table 3

Relative Hazard of Cardiovascular Disease After Adjustment for an Unmeasured Confounder in the Prevalent vs Incident Diabetes Cohort*

Prevalence of unmeasured confounder in prevalent cohort	Prevalence of unmeasured confounder in incident cohort	HR for CVD associated with unmeasured confounder	Adjusted HR† (2.24‡)	95% CI (2.08 to 2.40‡)
0.2	0.1	2.0	1.94	1.80 to 2.08
0.4	0.1	2.0	1.60	1.49 to 1.71
0.6	0.1	2.0	1.40	1.30 to 1.50
0.8	0.1	2.0	1.26	1.17 to 1.35
0.9	0.1	2.0	1.21	1.12 to 1.30
0.2	0.1	3.0	1.80	1.67 to 1.93
0.4	0.1	3.0	1.40	1.30 to 1.50
0.6	0.1	3.0	1.19	1.11 to 1.28
0.8	0.1	3.0	1.06	0.98 to 1.14
0.9	0.1	3.0	1.01	0.94 to 1.08

CVD = cardiovascular disease, HR = hazard ratio, CI = confidence interval

*A sensitivity analysis was performed to assess the association between diabetes duration and risk of CVD after adjustment for a hypothetical unmeasured confounder by varying the prevalence of the unmeasured confounder (0.2–0.9) in the prevalent diabetes cohort along with the relative hazard for CVD associated with the unmeasured confounder (HR = 2.0–3.0).

†Adjusted for age, sex, smoking (ever vs never), calendar year, and histories of hyperlipidemia, hypertension, and obesity (BMI ≥ 30).

‡Observed HR [95% CI] of CVD for the prevalent relative to incident diabetes cohort