2018

Validation of electronic medical data: Identifying diabetes prevalence in general practice

Joan Henderson
University of Sydney

Stephen Barnett
University of Wollongong, sbarnett@uow.edu.au

Abhijeet Ghosh
Coordinare Ltd,, ag477@uowmail.edu.au

Allan J. Pollack
University of Sydney

Adam J. Hodgkins
University of Wollongong, adamh@uow.edu.au

See next page for additional authors

Publication Details

Henderson, J., Barnett, S., Ghosh, A., Pollack, A. J., Hodgkins, A., Win, K., Miller, G. C. & Bonney, A. (2019). Validation of electronic medical data: Identifying diabetes prevalence in general practice. Health Information Management Journal, 48 (1), 3-11.

Validation of electronic medical data: Identifying diabetes prevalence in general practice

Abstract

Background: Electronic medical records are increasingly used for research with limited external validation of their data. Objective: This study investigates the validity of electronic medical data (EMD) for estimating diabetes prevalence in general practitioner (GP) patients by comparing EMD with national Bettering the Evaluation and Care of Health (BEACH) data. Method: A "decision tree" was created using inclusion/ exclusion of pre-agreed variables to determine the prob-ability of diabetes in absence of diagnostic label, including diagnoses (coded/free-text diabetes, polycystic ovarian syn-drome, impaired glucose tolerance, impaired fasting glucose), diabetic annual cycle of care (DACC), hemoglobin (HbA1 >6.5%, and prescription (metformin, other diabetes medications). Via SQL query, cases were identified in EMD of five Illawarra and Southern Practice Network practices (30,007 active patients; from 2 years to January 2015). Patient-based Supplementary Analysis of Nominated Data (SAND) sub-studies from BEACH investigating diabetes prevalence (1172 GPs; 35,162 patients; November 2012 to February 2015) were comparison data. SAND results were adjusted for number of GP encounters per year, per patient, and then age-sex standardised to match age-sex distribution of EMD patients. Cluster-adjusted 95% confidence intervals (CIs) were calculated for both datasets. Results: EMD diabetes prevalence (T1 and/or T2) was 6.5% (95% CI: 4.1-8.9). Following age-sex standardisation, SAND prevalence, not significantly different, was 6.7% (95% CI: 6.3-7.1). Extracting only coded diagnosis missed 13.0% of probable cases, subsequently identified through the presence of metformin/other diabetes medications medications (*without other indicator variables; 6.1%), free-text diabetes label (3.8%), HbA1c result* (1.6%), DACC* (1.3%), and diabetes medications* (0.2%). Discussion: While complex, proxy variables can improve usefulness of EMD for research. Without their consideration, EMD results should be interpreted with caution. Conclusion: Enforceable, transparent data linkages in EMRs would resolve many problems with identification of diagnoses. Ongoing data quality improvement remains essential.

Keywords

data:, electronic, medical, validation, general, prevalence, diabetes, practice, identifying

Publication Details

Henderson, J., Barnett, S., Ghosh, A., Pollack, A. J., Hodgkins, A., Win, K., Miller, G. C. & Bonney, A. (2019). Validation of electronic medical data: Identifying diabetes prevalence in general practice. Health Information Management Journal, 48 (1), 3-11.

Authors

Joan Henderson, Stephen Barnett, Abhijeet Ghosh, Allan J. Pollack, Adam J. Hodgkins, Khin Than Win, Graeme C. Miller, and Andrew D. Bonney

Research article

HIMJ

Health Information Management Journal
1–9
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1833358318798123
journals.sagepub.com/home/himj
SAGE

Validation of electronic medical data: Identifying diabetes prevalence in general practice

Joan Henderson, *BAppSC (HIM) Hons, PhD*[1],
Stephen Barnett, *BMed, PhD, DCH, MRCGP, FRACGP*[2],
Abhijeet Ghosh, *MBBS, GDIP (MedSonog), MSc (PopHealth)*[3],
Allan J Pollack, *MBBS(Hons), MBiomedE, FRACS (Orth), MPH(PP)*[1],
Adam Hodgkins, *BMED, DipPead, FRAACGP*[2],
Khin Than Win, *MBBS, PhD, DCS, IDCS, MS*[2],
Graeme C Miller, *MBBS, PhD, FRACGP*[1],
Andrew Bonney, *MBBS, MFM(Clin), PhD, FRACGP*[2]

Abstract

Background: Electronic medical records are increasingly used for research with limited external validation of their data. **Objective:** This study investigates the validity of electronic medical data (EMD) for estimating diabetes prevalence in general practitioner (GP) patients by comparing EMD with national Bettering the Evaluation and Care of Health (BEACH) data. **Method:** A "decision tree" was created using inclusion/exclusion of pre-agreed variables to determine the probability of diabetes in absence of diagnostic label, including diagnoses (coded/free-text diabetes, polycystic ovarian syndrome, impaired glucose tolerance, impaired fasting glucose), diabetic annual cycle of care (DACC), hemoglobin (HbA1c) > 6.5%, and prescription (metformin, other diabetes medications). Via SQL query, cases were identified in EMD of five Illawarra and Southern Practice Network practices (30,007 active patients; from 2 years to January 2015). Patient-based Supplementary Analysis of Nominated Data (SAND) sub-studies from BEACH investigating diabetes prevalence (1172 GPs; 35,162 patients; November 2012 to February 2015) were comparison data. SAND results were adjusted for number of GP encounters per year, per patient, and then age–sex standardised to match age–sex distribution of EMD patients. Cluster-adjusted 95% confidence intervals (CIs) were calculated for both datasets. **Results:** EMD diabetes prevalence (T1 and/or T2) was 6.5% (95% CI: 4.1–8.9). Following age–sex standardisation, SAND prevalence, not significantly different, was 6.7% (95% CI: 6.3–7.1). Extracting only coded diagnosis missed 13.0% of probable cases, subsequently identified through the presence of metformin/other diabetes medications medications (*without other indicator variables; 6.1%), free-text diabetes label (3.8%), HbA1c result* (1.6%), DACC* (1.3%), and diabetes medications* (0.2%). **Discussion:** While complex, proxy variables can improve usefulness of EMD for research. Without their consideration, EMD results should be interpreted with caution. **Conclusion:** Enforceable, transparent data linkages in EMRs would resolve many problems with identification of diagnoses. Ongoing data quality improvement remains essential.

Keywords (MeSH)

electronic medical records; data quality; data accuracy; general practice; primary health care; health information management

Introduction

At many levels of the Australian health system, electronic medical records (EMRs) are being employed increasingly for research purposes, as they are considered a timely and cost-effective method of providing data compared to the traditional techniques employed for structured prospective research studies (Shephard et al., 2011). In the primary medical care setting, commercial organisations, academic

[1] The University of Sydney, Australia,
[2] University of Wollongong, Australia
[3] Coordinare Ltd, Australia

Accepted for publication August 10, 2018.

Corresponding author:
Joan Henderson, School of Public Health, The University of Sydney, Sydney, NSW 2006, Australia.
E-mail: joan.henderson@sydney.edu.au

institutions, and government-funded not-for-profit organisations have formed collaborations with general practices to receive extracted patient data for analyses, for a variety of purposes (Mazza et al., 2016; MedicalDirector Research & Data Analytics, 2017; MedicineInsight, 2016).

There are significant limitations that influence the usefulness of these data but these limitations are not commonly addressed. There are more than eight different software products in use by Australian general practices (Gordon et al., 2016), all with different structures, having been independently created with different database designs'. None have enforceable, transparent direct linkages between the diagnosed problems/conditions being managed and the medications/other managements provided for these conditions. Clinicians are able to link clinical problems to medications but this is entirely elective, not always recorded accurately and frequently left incomplete. There are no nationally agreed and implemented standards, so these products have unique structures, data elements, data definitions, and data labels. As they utilise disparate terminology and classification systems (if any), pooling of these incompatible data is extremely problematic (Gordon et al., 2016). The data extraction tools used to access these data also have limitations (Liaw et al., 2013).

While the common use of extracted EMR data has introduced an implied "acceptability" for research, the validity and reliability of data extracted using these proprietary software products and extraction tools are unknown. Some organisations acknowledge the limitations of unknown data completeness and accuracy (Mazza et al., 2016; Merrifield et al., 2017) but there is no published evidence of attempts to validate extracted data.

In 2015, members of the Illawarra and Southern Practice Network (ISPRN) commenced an investigation with the aim of validating their pooled electronic medical data (EMD) by comparing it with data from the Bettering the Evaluation and Care of Health (BEACH) program (Britt et al., 2014). The initial results showed some similarities in patient demographics (age and sex) and in prescribing distribution patterns (Barnett et al., 2017). As an extension of that work, the current article describes the comparison of prevalence estimates for diabetes based on these two datasets.

Diabetes was selected as the condition of interest because it is commonly managed by general practitioners (GPs). The Australian national BEACH study of general practice activity reported an estimated 9.5% of patients at GP encounters having type 2 diabetes and 0.9% type 1 diabetes (Britt et al., 2014). It is also significant in terms of cost burden. Recent estimates put the annual cost of type 2 diabetes in Australia at around AUD6 billion with type 1 diabetes costing AUD570 million per year (Shaw et al., 2012). Diabetic medications were the most studied pharmacological class in the last decade (Mamtani et al., 2014), and diabetes is the most studied condition using data obtained from patient EMRs (Dean et al., 2009).

Ideally, researching disease prevalence using EMD would involve analyses of coded diagnoses extracted from the designated "diagnosis" field of the patient's EMR (Geraldine et al., 2012; Hwang et al., 2013). In reality (as explained above), diagnoses are not always coded (free text may be employed) or entered in the designated field – or recorded at all, and therefore are missing in data extraction.

In the absence of diagnostic labels and codes, pharmacological data have also been used as a proxy for a diabetes diagnosis (Huber et al., 2014). Although the Anatomical Therapeutic Chemical (ATC) classification (World Health Organisation, 2009) groups together medications used for diabetes, these may be prescribed for other conditions, reducing their reliability as indicators of diabetes. For example, metformin (ATC Code A10BA02) is commonly prescribed for diabetes but can also be used to treat polycystic ovarian syndrome (PCOS) and impaired glucose tolerance (IGT). Diabetes prevalence may be over-estimated if it is assumed that all ATC code A10 medications are prescribed only for diabetes. Conversely, some patients with diabetes may be missed if they are treated with non-pharmacological measures alone.

A combining of the diagnostic label and prescription data has previously been used to ascertain the presence of diabetes (Calvert et al., 2007; Chiang et al., 2014; Hasvold et al., 2014; MedicineInsight, 2016). The World Health Organisation defines a glycated hemoglobin (HbA1c) level of 6.5% (48 mmol/mol) or greater as diagnostic of diabetes (World Health Organisation, 2011). Even in the absence of relevant diagnoses or prescriptions, an elevated HbA1c implies diabetes.

In Australia, costs of GP consultations are subsidised through the universal insurance scheme, Medicare. Via the Medicare Benefits Schedule (MBS) (Australian Government Department of Health, 2014), a specific rebate item number is available to GPs for performing diabetic annual cycle of care (DACC). This involves a review of diabetic management and treatment goals specifically for patients with diagnosed diabetes. While claims data involving this item number have previously been used to assess diabetes management (Comino et al., 2015), no previous studies have used these data to determine diabetes prevalence.

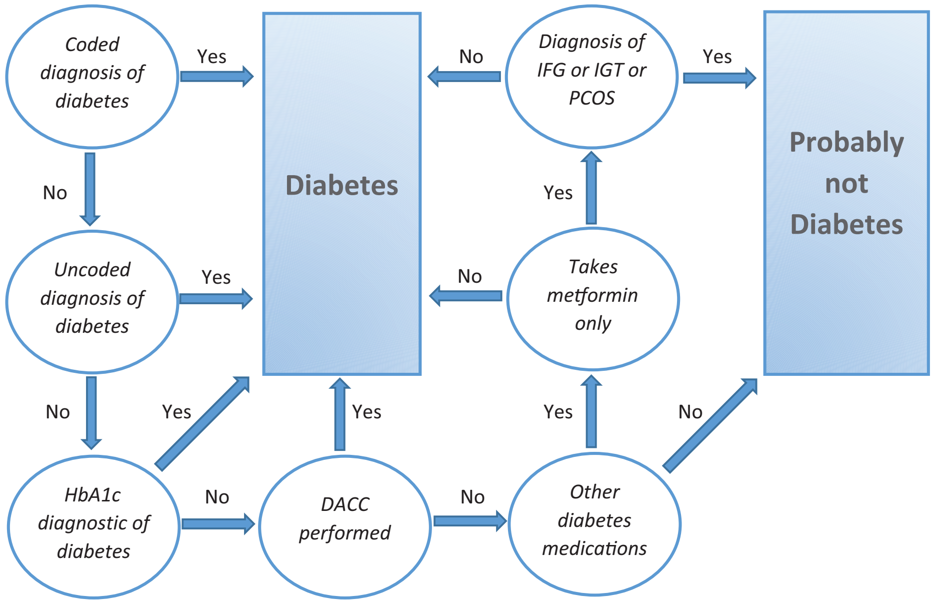
In Sweden, Rolandsson et al. (2012) demonstrated the validity of the diagnosis of diabetes based on the information in general practice medical records (including the coded diagnosis, pharmacological data, and records of HbA1c results) (Rolandsson et al., 2012). However, we have been unable to identify any previously published Australian studies that have attempted to validate data from EMRs in the determination of a diabetes diagnosis.

Aim of this study investigates the validity of EMD for estimating the prevalence of diabetes in general practice patients by comparing data extracted from EMRs with prevalence data from patient-based sub-studies of the University of Sydney's national BEACH program (Britt et al., 2014).

Methods

The EMD set came from the ISPRN based at the University of Wollongong. There were 40 general practices in the ISPRN network when the project commenced, and a convenience sample of six member practices (chosen because they varied in size and geographic location) was invited to

Box 1. The EMD decision tree. EMD: electronic medical data.


Coded diagnosis of diabetes
Yes
No
Diabetes
No
Diagnosis of IFG or IGT or PCOS
Yes
Probably not Diabetes
Yes
Uncoded diagnosis of diabetes
Yes
No
Takes metformin only
No
Yes
Yes
Yes
No
HbA1c diagnostic of diabetes
No
DACC performed
No
Other diabetes medications

participate. All used Best Practice™ clinical software, ensuring standard structure and terminology.

Researchers from both the University of Wollongong and the University of Sydney met via teleconference to plan the project method. Based on the literature review (above) and the clinicians' knowledge, the authors agreed on a set of variables for inclusion in a data extraction query, in addition to coded diagnoses, in case a coded diagnosis had not been recorded.

EMD set

An SQL query was created to extract de-identified patient demographic and medical data variables of interest: age; sex; diagnosis of diabetes, PCOS, IGT, or impaired fasting glucose (IFG; coded or uncoded, i.e. free text); diabetes cycle of care plan (DACC); recorded HbA1c value of $\geq$6.5% (48 mmol/mol); prescription of metformin; and/or prescription of other diabetes medications. The research team (seven clinicians (four with PhD, one with masters in clinical information systems and technology, and one with masters in biostatistics) and one health information manager (PhD in medicine (general practice)) determined that this group of variables was the best indicator of diabetes if the record lacked the diagnosis label and, accordingly, created a "decision tree" (Box 1). Therefore, in the absence of "diabetes" in a diagnosis label, the presence of diabetes in the patient was inferred based on the "decision tree" variables.

Best Practice software uses its own bespoke PYEFINCH coding system. Codes identifying diabetes, PCOS, IGT, or IFG were selected for the query, to allow subsequent exclusion of the last three if these had been incorrectly included through use of metformin or other diabetes medication. Metformin and other diabetes medications were identified by the presence in the EMR of a prescription for a medication classified to the ATC level A10 "Drugs used in Diabetes" (World Health Organisation Collaborating Centre of Drug Statistics Methodology, 2009). Previous DACC was determined by the presence of relevant MBS item numbers.

The query was designed and corrected as necessary to ensure that the required variables were extracted. The patient cohort consisted of active patients, defined as patients who had had three or more encounters in the previous 2 years (Royal Australian College of General Practitioners, 2014). For this study, active patients at any one of the participation practices were included from the 2 years prior to January 1, 2015. Encrypted data were sent from each practice to the investigators, by online secure transfer.

Comparison dataset: SAND sub-studies of the BEACH program

The data used in the comparison analysis for validating the prevalence estimate were collected through a series of Supplementary Analysis of Nominated Data (SAND) sub-studies of the BEACH program. BEACH was a continuous, national, cross-sectional survey of Australian general practice activity. The BEACH methods have proven statistical validity and reliability and have been described internationally as "the gold standard from consistent reporting". The methods are described in detail elsewhere (Britt et al., 2014), but in brief, each year approximately 1000 randomly sampled GPs were recruited. Each participant recorded details of 100 encounters with consenting, unidentified patients, on structured paper forms. Information was collected about the problem(s) managed for each patient, and the management(s) provided (directly linked) for those problems. Throughout the program, a series of SAND sub-studies utilised the GP as an "expert interviewer" to

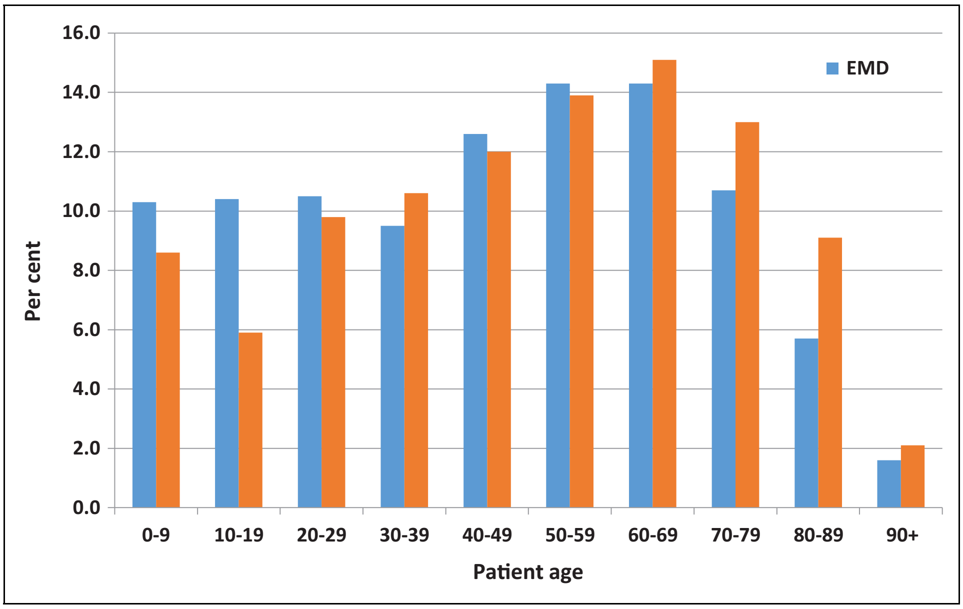


Figure 1. Age distribution of patients in EMD records and in SAND sub-studies. EMD: electronic medical data; SAND: Supplementary Analysis of Nominated Data.

record, in discussion with the patient, the aspects of patient health that may not have been managed at the recorded encounter. The SAND data are, therefore, patient-based rather than encounter-based.

The SAND sub-studies were selected as the best dataset for comparison because they were designed for active data collection for the purpose of research. Thus, the presence or absence of diabetes was the response recorded by the GP, not inferred by variables suggestive of diabetes. Between November 2012 and February 2015, 12 of the SAND sub-studies, each involving approximately 125 GPs, were used to investigate the prevalence of diagnosed chronic conditions for 30 of every 100 patients surveyed. GPs were asked to record for the patient at each encounter, all their diagnosed chronic conditions, including diabetes type 1 and/or diabetes type 2, based on their knowledge of the patient, the patient's notes, and the patient's knowledge of their own medical conditions. The number of times the patient had seen any GP in the previous 12 months was also recorded.

Statistical analyses

Using unadjusted SAND data, we estimated the prevalence of diabetes among patients sampled at encounters. As BEACH patients were sampled at GP encounters, the likelihood of being sampled is dependent on visit frequency. Therefore, frequent attenders (older patients, patients with one or multiple chronic conditions) are more likely to be sampled than those who attend less frequently. To allow fair comparison with EMD-derived results, the SAND results were adjusted for the reported number of GP encounters per year for each patient, allowing the creation of prevalence estimates among the attending population. This "active patient" sample was then age–sex standardised, by weighting the SAND data to match the age–sex distribution of EMD patients. Missing data were excluded.

Both the BEACH program and the EMD project used a cluster survey design (clustered around GP participants for BEACH and around general practices for EMD). BEACH analyses adjusted for the effect of cluster when 95% confidence intervals (CIs) were calculated around all point estimates, using survey procedures in SAS$^{®}$ 9.3 (Cary, North Carolina, USA, in 2012). Similarly, a cluster-adjusted 95% CI was calculated for the aggregated EMD diabetes prevalence estimate, using Microsoft Excel (v. 2013; Microsoft Corporation, Redmond, Washington, USA). Differences between EMD and adjusted SAND estimates were considered statistically significant ($p < 0.05$) if the CIs did not overlap.

Ethics

The research study was approved by the Human Research Ethics Committee (HREC) of the University of Wollongong/Illawarra Shoalhaven Local Health District (HREC approval number HE13/484, November 21, 2013). Signed consent was obtained from all practice principals involved prior to data extraction, and downloaded patient data were de-identified prior to extraction. The BEACH program and all SAND sub-studies have ethics approval and oversight by the HREC of the University of Sydney (HREC approval number 2012-130, valid to Match 31, 2018).

Results

Five of the six ISPRN practices agreed to participate. Data from these five practices resulted in a sample of 30,007 patients. The "date of birth" field was not completed for 39 patients (0.13%). For the remainder, the median age was 47 years (interquartile range (IQR) 23–65). Persons aged 65 years and over accounted for 25.6% of the sample (Figure 1) and females for 56.4%.

In the BEACH study, of the 1500 GPs who were recruited and sent recording forms containing the sub-

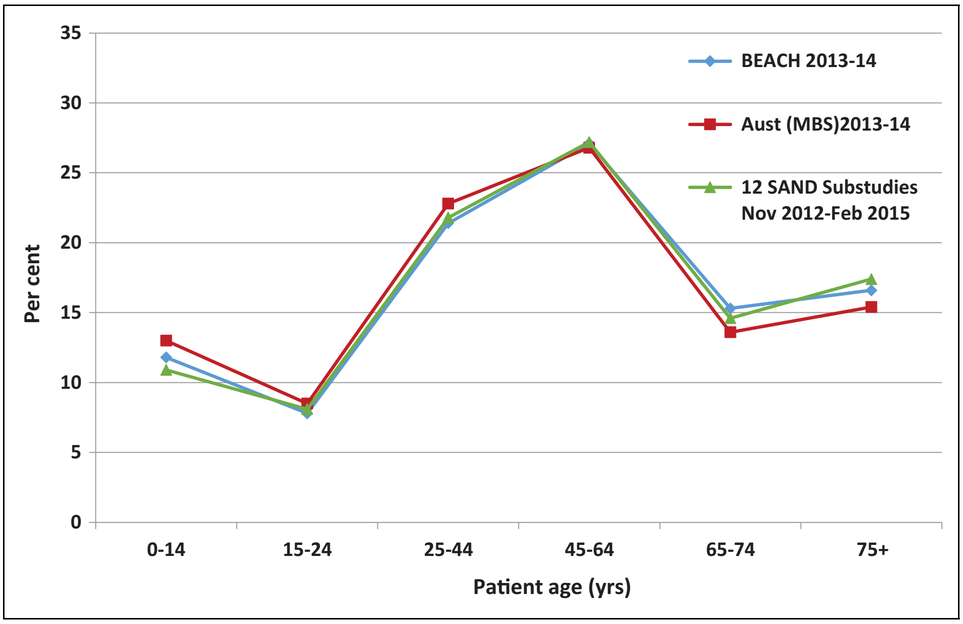


Figure 2. Age distribution of patients: At BEACH encounters, at claimed MBS GP consult service items, and at encounters in the SAND sub-studies. BEACH: Bettering the Evaluation and Care of Health; MBS: Medicare Benefits Schedule; GP: general practitioner; SAND: Supplementary Analysis of Nominated Data.

study questions about diagnosed chronic disease and number of GP visits, 1172 returned completed questionnaires (78.1%) for 35,162 GP-patient encounters. The "date of birth" field was incomplete for 267 patients (0.76%). The patient median age was 52 years (IQR 31–69), 32.5% being aged 65 years and over, and 59.7% being female. The SAND sub-studies were chosen for comparison because they are nationally representative. Figure 2 shows the age distribution of patients: at all BEACH encounters for 2013–2014, for whom MBS GP consultation service items were claimed through Medicare in 2013–2014, and at the BEACH encounters where the SAND data were collected. The greatest disparity in any age group (compared with the Medicare distribution) is less than 2% points.

At the first step of the EMD diabetes prevalence model (decision tree), there were 1700 patients with a coded diagnosis of diabetes, giving a prevalence of 5.7% based on this variable alone. When all proxy variables were assessed, further 253 patients were identified as probably having diabetes, increasing the total number to 1953 and the prevalence to 6.5% (95% CI: 4.1–8.9; types 1 and/or 2 combined; Figure 3).

Each proxy variable contributed different proportions to the 253 extra cases identified (13.0% of 1953). The free-text (uncoded) search for diabetes accounted for 29.6% of the 253 additional cases, but the major contributor was the prescription of metformin (47.0% of 253; Figure 4). Using this process did not allow differentiation between types 1 and type 2 diabetes.

From the SAND data, we could estimate the prevalence of types 1 and 2 diabetes separately, but for comparison with the EMD dataset, these were combined. The combined prevalence among sampled patients was 10.3% (95% CI: 9.9–10.6), active patients (i.e. adjusted for number of visits) was 5.6% (5.3–5.8), and "age-sex standardised" active patients was 6.7% (6.3–7.1). This 6.7% did not significantly differ from the EMD sample estimate above (6.5%; 4.1–8.9; $p = 0.8$, Satterthwaite method, equivalent to $t = 0.21$, $df = 33{,}123$).

Discussion

This study tested the validity of EMD to estimate the prevalence of diabetes in general practice patients by comparing data extracted from EMRs with prevalence data gathered as part of the national BEACH program. The findings add to the ongoing debate about the usefulness of EMD data for research in its current state (Hersh et al., 2013). By recognising that GP data are often incomplete, and by employing proxy variables where coded diagnoses were not recorded, the EMD has produced a prevalence estimate for diabetes that is not significantly different to that yielded by a dataset of proven validity and reliability.

The fact that approximately 13% of diabetes cases were not coded when a coding system was available in the software supports the need to foster ongoing efforts to improve data quality. However, by compensating with proxy measures such as medications, free-text searches of other fields, item numbers, and the consideration/exclusion of other diagnoses, the usefulness of EMD can be improved, given access to good SQL skills or designed-for-purpose data extraction tools.

The results of this study also highlight that not all proxies are equal. Individual variables, and the order in which they are queried, add differing numbers of cases. For example, the simple addition of "uncoded" free-text diagnoses of diabetes accounted for 29.6% of the additional cases. The major contributor to additional likely cases (47.0%) was the subgroup of patients taking metformin

Figure 3. Diabetes prevalence using the EMD Decision Tree.

Indicator	*n* *(30,007)*	Cumulative *n*	Cumulative %
Coded diagnosis of diabetes ONLY	1700	1700	**5.67**
NO coded diagnosis **YES uncoded diagnosis of diabetes**	75	1775	**5.92**
NO coded / uncoded diagnosis of diabetes NO diagnosis of PCOS / IFG / IGT NO diagnosis of **YES HbA1c**	31	1806	**6.02**
NO coded / uncoded diagnosis of diabetes NO diagnosis of PCOS / IFG / IGT NO HbA1c **YES DACC**	25	1831	**6.10**
NO coded / uncoded diagnosis of diabetes NO diagnosis of PCOS / IFG / IGT NO HbA1c NO DACC **YES EITHER metformin OR other drugs used in diabetes**	119	1950	**6.50**
NO coded / uncoded diagnosis of diabetes **YES EITHER diagnosis of PCOS OR Diagnosis of IFG/IGT** NO HbA1c NO DACC NO metformin **YES other drugs used in diabetes**	3	1953	**6.51**
95% CI			6.23–6.79

EMD: electronic medical data; PCOS: polycystic ovarian syndrome; IFG: impaired fasting glucose; IGT: impaired glucose tolerance; HbA1c: hemoglobin; CI: confidence interval; DACC: diabetic annual cycle of care.

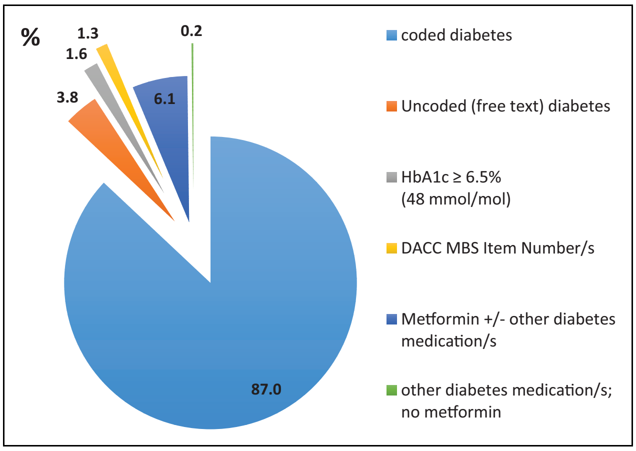


Figure 4. Contribution of coded diabetes and of proxy variables (i.e. all variables except "coded diabetes") to diagnosis count.

who had no (coded or uncoded) diagnoses of diabetes, PCOS, or IFG and no HbA1c result. Metformin alone as a search term may be associated with conditions other than diabetes (such as PCOS, impaired fasting glycaemia, or weight gain) and may thus lead to false positives.

Our "step-wise" use of metformin minimises the risk of inaccuracy, by excluding potential false positives before the metformin criterion is applied. Despite this approach, many EMD studies may require manual review of at least a representative sample of records, to further clarify the accuracy of the results. This is a consideration for future studies. In contrast, the use of non-metformin oral hypoglycaemic agents and of insulin is strongly correlated with diabetes, and the number of additional cases is very likely to be accurate.

The design of this method involved clinicians and academics who understand how diabetes is managed in general practice, which allowed the creation of a tool most likely to reduce false negatives. The designers felt confident that the proxies selected for an investigation of diabetes should result in very few missed cases. These proxies, however, were not able to identify type 1 as distinct from type 2, and possibly included gestational diabetes or diabetes insipidus, where either the term "diabetes" or the proxies were searched as free text. Unless there was some clear indicator that the diabetes (actual or inferred) was other than types 1 or 2, it will not have been excluded. Another consideration in using medications as proxy variables to distinguish between diabetes types is the degree to which the use of these medications overlaps – metformin is used to treat type 2 diabetes, but may also be used for gestational diabetes. Insulin can no longer be considered a point of differentiation between types 1 and 2 as the progression of type 2 diabetes increasingly involves the eventual requirement for insulin (Diabetes Australia, 2015). These diagnosed conditions were clearly differentiated using the traditional research methods employed in the SAND sub-studies of BEACH.

While undertaken in the Australian setting, this work has international implications. Studies from the United Kingdom (Calvert et al., 2007; Majeed et al., 2008; Muller, 2014; Shephard et al., 2011) and the United States (Hersh et al., 2013; Kamal et al., 2014) have reported advantages of using EMR data for research, but in all cases have also acknowledged the weaknesses of these data and the impact on data quality. There is evidence that different types of variables may have different GP completeness-of-recording rates. This may affect the capacity to produce prevalence data when investigating other diseases or population sub-groups. For example, demographic variables such as age and sex have a high completion rate in EMRs. It is obvious when a "date of birth" or "age" is missing as all patients have one. For both datasets in this study, the proportion of missing data for the recorded "date of birth" was clear – 267 patients (0.76%) in the SAND sample and 39 patients (0.13%) in the EMD sample. However, other variables such as "smoking status" may be considered differently. If the GP considers the patient's smoking status to have a clinical bearing only if the patient does smoke, this may influence the decision to complete this variable for nonsmokers – which decreases the variable's reliability. In a previous comparison of the EMD with BEACH data, there was a significant difference ($p = 0.03$) in proportion of current smokers between the two datasets, possibly influenced by the high proportion of missing "smoking status" data (17.6%) in the EMD sample compared with BEACH data (1.8%) (Barnett et al., 2017).

Other researchers have commented on the differing "value" of variables and its effect on completeness. Mazza et al. (2016) suggested that "because the primary use of the data is to inform the clinical care of a patient, only data that serve a clinical purpose have a high degree of validity and reliability. Therefore, items such as Aboriginality are not necessarily well recorded."

A number of authors in the United States and the United Kingdom (e.g. Kamal et al., 2014; Muller, 2014) have acknowledged that some variables are more complete than others and many factors will influence this aspect of data quality. Muller (2014) suggests that the health system of a country may influence prescribing behaviour as in some cases, a medication that is available without a prescription may be advised for some patients and prescribed for others, depending on their income and circumstances. In the US study by Kamal et al. (2014), the researchers reported "incomplete data sets for certain patient characteristics (race, marital status, and employment" and "the EMR data did not contain information related to lifestyle measures such as exercise and diet." This latter exclusion will also have had an impact on the current study as any patients who did not have a diabetes label acknowledging their diagnosis may have been missed unless other proxy variables captured them. A further consideration when selecting proxy variables is the approximately 24% of Australian general practice patients with type 2 diabetes who have their condition managed with diet and exercise only (Family Medicine Research Centre, University of Sydney, 2012). These patients would not be included in any prevalence estimate reliant on medication alone as a proxy for identification.

Although there is a consensus that reliable data are essential – for clinician and patient education, for patient safety, for financial planning and resource allocation – what should be acknowledged is that EMRs were designed for the clinician to keep track of their patient's care, not as a research tool. GPs are time poor, and the variables being presented to them for recording patient information in currently available clinical software programs are numerous. The computer already imposes on the time a GP might otherwise give to the patient (Dowell et al., 2013; Haywood et al., 2015; Pearce et al., 2011), so there is a selection process occurring that meets the primary objective, that is, the care of the patient. Many GPs may not think it important to "patient care" to record the indication associated with a medication or a test in a specified area – they know why the script was given or the test ordered – and this absence limits the capacity of EMRs to produce a reliable prevalence estimate. Where no diagnosis is recorded and a medication or test could be for a number of possible indications, the method developed for the EMD sample in this diabetes investigation would be far less reliable when applied to other conditions. Data linkages in EMRs would resolve many of these problems with identification of diagnoses. At present, all research based on data extraction from EMRs faces the same problem. Assessing the extent of missing data would allow the measurement of data quality but for many variables, the proportion of missing data is not known and the validity and reliability of any calculations using an unknown denominator cannot be estimated.

"The completeness and accuracy of data entry relies mainly on the enthusiasm of family practitioners" (Majeed et al., 2008). This statement is highly appropriate to this investigation. While these results have provided some evidence of the proportion of diabetes being missed with current extractions that rely on coded diagnosis fields only (13.0%), it should be acknowledged that the five participating practices contributing to the EMD set consist of GPs who are highly "data conscious" and because of this awareness, and their ongoing quality improvement activities, EMD from a different set of practices may have produced a different result.

This work is, therefore, both enlightening and valuable. It is the first attempt by any Australian researchers to validate the data from their patient EMRs in this manner, to assess its fitness as a research tool. The results have shown that while it may be a more complex approach, the use of proxy variables can improve the usefulness of EMD as a research tool, in some circumstances. The result was promising with diabetes and the researchers had planned a series of investigations on other chronic diseases using the BEACH and SAND sub-studies for validation. The BEACH datasets, while not electronic, have been proven repeatedly to be valid, reliable representations of general practice activity and are, therefore, highly appropriate datasets to employ for this purpose. Given the BEACH project has now ceased, the work undertaken for this study is very timely – it will not be easy to find a comparable dataset of its calibre in the future.

Acknowledgements

The authors would like to thank the Illawarra and Shoalhaven Medicare Local for funding and support; the members of the EMD Group, ISPRN, and particularly Alyssa Horgan, for administrative support; Pippa Burns and Jack Bird for research support; Christopher Harrison for statistical advice; Prof Helena Britt for proofreading and comment; and Prof Don Iverson for his vision and encouragement. We thank the GPs who participated in BEACH for the generous donation of their time.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The BEACH program was funded by a consortium of pharmaceutical manufacturers, Australian government agencies, and government-funded quality of care agencies. Each group contributed funding through a research agreement with the University of Sydney which provided complete research autonomy for the team conducting the BEACH Program. Funders provided input into the research design development through an Advisory Board; however, the final decisions regarding research design, data collection and analysis, and reporting of findings remain with the principal investigators under the ethical supervision of the University. During the period involved in this study, BEACH was supported by the Australian Government Department of Health, AstraZeneca Pty Ltd (Australia), Novartis Pharmaceuticals Australia Pty Ltd, bioCSL (Australia) Pty Ltd, Merck, Sharp & Dohme (Australia) Pty Ltd, Pfizer Australia, the National Prescribing Service, and the Australian Government Department of Veterans' Affairs. BEACH was endorsed by the AMA and the RACGP.

ORCID iD

Andrew Bonney http://orcid.org/0000-0003-2477-1646

References

Australian Government Department of Health (2014) Medicare benefits schedule book. Available at: http://www.health.gov.au/internet/mbsonline/publishing.nsf/Content/432EE55FAB58E5C4CA257D6B001AFB8A/$File/201411-MBS.pdf (accessed 2 February 2016).

Barnett S, Henderson J, Hodgkins A, et al. (2017) A valuable approach to the use of electronic medical data in primary care research: panning for gold. *Health Information Management Journal* 46(2): 51–57.

Britt H, Miller GC, Henderson J, et al. (2014) *General Practice Activity in Australia 2013-14*. Sydney: Sydney University Press.

Calvert MJ, McManus RJ and Freemantle N (2007) Management of type 2 diabetes with multiple oral hypoglycaemic agents or insulin in primary care: retrospective cohort study. *British Journal of General Practice* 57(539): 455–460.

Chiang HH, Tseng FY, Wang CY, et al. (2014) All-cause mortality in patients with type 2 diabetes in association with achieved hemoglobin A(1c), systolic blood pressure, and low-density lipoprotein cholesterol levels. *Plos One* 9: e109501. Epub ahead of print 27 October 2014. DOI: 10.1371/journal.pone.0109501.

Comino EJ, Islam MF, Tran DT, et al. (2015) Association of processes of primary care and hospitalisation for people with diabetes: a record linkage study. *Diabetes Research and Clinical Practice* 108(2): 296–305.

Dean BB, Lam J, Natolk JL, et al. (2009) Review: use of electronic medical records for health outcomes research: a literature review. *Medical Care Research and Review* 66(6): 611–638.

Diabetes Australia (2015) Type 2 diabetes. Available at: https://www.diabetesaustralia.com.au/type-2-diabetes (accessed 2nd May 2018).

Dowell A, Stubbe M, Scott-Dowell K, et al. (2013) Talking with the alien: Interaction with computers in the GP consultation. *Australian Journal of Primary Health* 19: 275–282.

Family Medicine Research Centre, University of Sydney (2012) *SAND Abstract No. 185 from the BEACH Program: Diabetes Management and Self-monitoring in General Practice Patients*. Sydney: FMRC University of Sydney 2012. ISSN 1444-9072. Available at: http://sydney.edu.au/medicine/fmrc/publications/sand-abstracts/185-Diabetes_management_and_self-management.pdf (accessed 2nd May 2018).

Geraldine N, Marc A, Carla T, et al. (2012) Relation between diabetes, metformin treatment and the occurrence of malignancies in a Belgian primary care setting. *Diabetes Research and Clinical Practice* 97(2): 331–336.

Gordon J, Miller G, Britt H, et al. (2016) Reality check – reliable national data from general practice electronic health records. Available at: http://ahha.asn.au/publication/issue-briefs/deeble-institute-issues-brief-no-18-reality-check-reliable-national-data (accessed 3 March 2017).

Hasvold LP, Bodegard J, Thuresson M, et al. (2014) Diabetes and CVD risk during angiotensin-converting enzyme inhibitor or angiotensin II receptor blocker treatment in hypertension: a study of 15,990 patients. *Journal of Human Hypertension* 28(11): 663–669.

Haywood J, Buckingham S, Thomson F, et al. (2015) 'How long does it take?' A mixed-method evaluation of computer-related work in GP consultations. *Journal of Innovation in Health Informatics* 22(4): 409–425.

Hersh WR, Weiner MG, Embi PJ, et al. (2013) Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care* 51(8 suppl 3): S30–S37.

Huber CA, Diem P, Schwenkglenks M, et al. (2014) Estimating the prevalence of comorbid conditions and their effect on health care costs in patients with diabetes mellitus in Switzerland. *Diabetes Metabolic Syndrome and Obesity* 7: 455–465.

Hwang A, Narayan V and Yang YX (2013) Type 2 diabetes mellitus and survival in pancreatic adenocarcinoma: a retrospective cohort study. *Cancer* 119(2): 404–410.

Kamal KM, Chopra I, Elliott JP, et al. (2014) Use of electronic medical records for clinical research in the management of type 2 diabetes. *Research in Social and Administrative Pharmacy* 10: 877–884.

Liaw ST, Taggart J, Yu H, et al. (2013) Data extraction from electronic health records – existing tools may be unreliable and potentially unsafe. *Australian Family Physician* 42(11): 820–823.

Majeed A, Carr J and Sheikh A (2008) Accuracy and completeness of electronic patient records in primary care. *Family Practice* 25(4): 213–214.

Mamtani R, Haynes K, Finkelman BS, et al. (2014) Distinguishing incident and prevalent diabetes in an electronic medical records database. *Pharmacoepidemiology Drug Safety* 23(2): 111–118.

Mazza D, Pearce C, Turner LR, et al. (2016) The Melbourne East Monash general practice database (MAGNET): using data from computerised medical records to create a platform for primary care and health services research. *Journal of Innovation in Health Informatics* 23(2): 523–528.

MedicalDirector. Research & Data Analytics (2017) Available at: http://medicaldirector.com/Pi+Ch/Pi+Ch+Research+And+Data+Analytics (accessed 1 March 2017).

MedicineInsight (2016) *MedicineInsight Data Book Version 1.2. NPS MedicineWise*, Sydney. Available at: http://www.nps.org.au/__data/assets/pdf_file/0019/324415/MedicineInsight-Databook-v1.2.pdf (accessed 3 March 2017).

Merrifield A, Gilles MB, Belcher J, et al. (2017) Why are antidepressants prescribed in Australian general practice? Available at: http://www.phcris.org.au/phplib/filedownload.php?file=/elib/lib/downloaded_files/conference/presentations/8147_conf_abstract_.pdf (accessed 5 March 2017).

Muller S (2014) Electronic medical records: the way forward for primary care research? *Family Practice* 31(3): 127–129.

Pearce C, Arnold M, Phillips C, et al. (2011) The patient and the computer in the primary care consultation. *Journal of American Medical Informatics Association* 18: 138–142.

Rolandsson O, Norberg M, Nystrom L, et al. (2012) How to diagnose and classify diabetes in primary health care: lessons learned from the diabetes register in Northern Sweden (DiabNorth). *Scandinavian Journal of Primary Health Care* 30(2): 81–87.

Royal Australian College of General Practitioners (2014) *Standards for general practices*. 4th ed. *Glossary of Terms*. Available at: http://www.racgp.org.au/your-practice/standards/standards4thedition/appendices/glossary-of-terms/ (accessed 11 May 2017).

Shaw J and Tanamas S (eds.), Baker IDI Heart and Diabetes Institute, et al. (2012) Diabetes: the silent pandemic and its impact on Australia. Available at: https://static.diabetesaustralia.com.au/s/fileassets/diabetes-australia/e7282521-472b-4313-b18e-be84c3d5d907.pdf (accessed 5 November 2017).

Shephard E, Stapley S and Hamilton W (2011) The use of electronic databases in primary care research. *Family Practice* 28(4): 352–354.

World Health Organisation (2011). *Use of Glycated Haemoglobin (HbA1c) in Diagnosis of Diabetes Mellitus: Abbreviated Report of a WHO consultation*. Geneva: WHO.

World Health Organisation Collaborating Centre of Drug Statistics Methodology (2009) *Anatomical Therapeutic Chemical (ATC) Classification Index with Defined Daily Doses (DDDs)*. Oslo: WHO.