Springer

Dear Author

Here are the proofs of your article.

- You can submit your corrections **online, by email** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- For **fax** submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the article number, and your name when sending your response via e-mail, or fax.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during typesetting and insert your answers/corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor.
- If we do not receive your corrections **within 4 days**, we will send you a reminder.

Please note

Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL:

`http://dx.doi.org/10.1007/s00125-017-4518-6`

If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information, go to: http://www.springerlink.com.

The **printed version** will follow in a forthcoming issue.

Fax to: +44 (0)117 4147887

Diabetologia Editorial Office
(diabetologia-j@bristol.ac.uk)

From: **Diabetologia DOI 10.1007/s00125-017-4518-6**
Re: Promises and pitfalls of electronic health record analysis

Authors: **Farmer · Mathur · Bhaskaran · Eastwood · Chaturvedi · Smeeth**

Permission to publish

Dear Editorial Office,

I have checked the proofs of my article and

- ☐ I have **no corrections.** The article is ready to be published without changes.
- ☐ I have **a few corrections.** I am enclosing the following pages:
- ☐ I have made **many corrections.** Enclosed is the **complete article.**

Metadata of the article that will be visualized in OnlineFirst

1	Article Title	**Promises and pitfalls of electronic health record analysis**	
2	Article Sub- Title		
3	Article Copyright - Year	**The Author(s) 2017** **(This will be the copyright line in the final PDF)**	
4	Journal Name	Diabetologia	
5	Corresponding Author	Family Name	**Farmer**
6		Particle	
7		Given Name	**Ruth**
8		Suffix	
9		Organization	London School of Hygiene and Tropical Medicine
10		Division	Department of Non-communicable Disease Epidemiology
11		Address	Keppel Street, London WC1E 7HT
12		e-mail	ruth.farmer@lshtm.ac.uk
13	Author	Family Name	**Mathur**
14		Particle	
15		Given Name	**Rohini**
16		Suffix	
17		Organization	London School of Hygiene and Tropical Medicine
18		Division	Department of Non-communicable Disease Epidemiology
19		Address	Keppel Street, London WC1E 7HT
20		e-mail	
21	Author	Family Name	**Bhaskaran**
22		Particle	
23		Given Name	**Krishnan**
24		Suffix	
25		Organization	London School of Hygiene and Tropical Medicine
26		Division	Department of Non-communicable Disease Epidemiology
27		Address	Keppel Street, London WC1E 7HT
28		e-mail	
29	Author	Family Name	**Eastwood**

30		Particle	
31		Given Name	**Sophie**
32		Suffix	
33		Organization	Institute for Cardiovascular Sciences, University College London
34		Division	
35		Address	London
36		e-mail	
37		Family Name	**Chaturvedi**
38		Particle	
39		Given Name	**Nishi**
40		Suffix	
41	Author	Organization	Institute for Cardiovascular Sciences, University College London
42		Division	
43		Address	London
44		e-mail	
45		Family Name	**Smeeth**
46		Particle	
47		Given Name	**Liam**
48		Suffix	
49	Author	Organization	London School of Hygiene and Tropical Medicine
50		Division	Department of Non-communicable Disease Epidemiology
51		Address	Keppel Street, London WC1E 7HT
52		e-mail	

53		Received	23 June 2017
54	Schedule	Revised	
55		Accepted	24 October 2017

56 Abstract

Routinely collected electronic health records (EHRs) are increasingly used for research. With their use comes the opportunity for large-scale, high-quality studies that can address questions not easily answered by randomised clinical trials or classical cohort studies involving bespoke data collection. However, the use of EHRs generates challenges in terms of ensuring methodological rigour, a potential problem when studying complex chronic diseases such as diabetes. This review describes the promises and potential of EHRs in the context of diabetes research and outline key areas of caution with examples. We consider the difficulties in identifying and classifying diabetes patients, in distinguishing between prevalent and incident cases and in dealing with the

		complexities of diabetes progression and treatment. We also discuss the dangers of introducing time-related biases and describe the problems of inconsistent data recording, missing data and confounding. Throughout, we provide practical recommendations for good practice in conducting EHR studies and interpreting their results.
57	Keywords separated by ' - '	Diabetes - Electronic health records - Epidemiology - Observational studies - Primary care - Review - Secondary care
58	Foot note information	Ruth Farmer and Rohini Mathur contributed equally to this work.

Diabetologia
https://doi.org/10.1007/s00125-017-4518-6

REVIEW

Promises and pitfalls of electronic health record analysis

Ruth Farmer[1] · Rohini Mathur[1] · Krishnan Bhaskaran[1] · Sophie Eastwood[2] · Nishi Chaturvedi[2] · Liam Smeeth[1]

Received: 23 June 2017 / Accepted: 24 October 2017

Abstract
Routinely collected electronic health records (EHRs) are increasingly used for research. With their use comes the opportunity for large-scale, high-quality studies that can address questions not easily answered by randomised clinical trials or classical cohort studies involving bespoke data collection. However, the use of EHRs generates challenges in terms of ensuring methodological rigour, a potential problem when studying complex chronic diseases such as diabetes. This review describes the promises and potential of EHRs in the context of diabetes research and outline key areas of caution with examples. We consider the difficulties in identifying and classifying diabetes patients, in distinguishing between prevalent and incident cases and in dealing with the complexities of diabetes progression and treatment. We also discuss the dangers of introducing time-related biases and describe the problems of inconsistent data recording, missing data and confounding. Throughout, we provide practical recommendations for good practice in conducting EHR studies and interpreting their results.

Keywords Diabetes · Electronic health records · Epidemiology · Observational studies · Primary care · Review · Secondary care

Abbreviations
CKD Chronic kidney disease
CVD Cardiovascular disease
EHR Electronic health record
GP General practitioner/General practice

Introduction

A greater understanding of the changing patterns of treatment, patient demographics, risk factors and disease burden is vital to inform clinical care and public health policy in diabetes. RCTs are key but will not answer all questions as they have several limitations: (1) they often have insufficient power and length of follow-up to examine clinical endpoints; (2) aspects of patient behaviour and clinical care are likely to differ in trials compared with real-world settings and (3) important groups, such as women of childbearing age, individuals with multimorbidities and ethnic minorities, may be under-represented in clinical trial populations [1–3]. Similarly, classical cohort studies involving bespoke data collection are expensive and time consuming and rarely have long-term follow-up for participants beyond the initial study period.

The use of electronic health records (EHRs) for research allows us to overcome many of these limitations and address important scientific questions. Post marketing and surveillance studies using EHRs are key for speeding up access to new drugs [4]. Recognising this, the ADA recently endorsed the use of evidence from high-quality observational studies to aid therapeutic decision making [5, 6]. In recent years, the use of EHRs for research has grown tremendously and the potential for observational studies using EHRs to generate valid estimates of causal associations is beginning to be explored. Though EHRs have the potential to produce high-quality research, major challenges exist. In this narrative review, we describe the promises and potential of EHRs, outline some key areas of caution and provide practical recommendations for using EHRs in the context of diabetes research.

Ruth Farmer and Rohini Mathur contributed equally to this work.

✉ Ruth Farmer
ruth.farmer@lshtm.ac.uk

[1] Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

[2] Institute for Cardiovascular Sciences, University College London, London, UK

The promise of EHR data

The term 'electronic health record' encompasses a wide variety of data sources including, but not limited to, routinely collected primary and secondary care records, disease-specific registries and health insurance claims databases (Table 1). Several key potential advantages of EHRs are outlined in the Text box.

EHRs are widely used to enable contemporary estimation of disease incidence or prevalence [13–15], study prospective associations between risk factors and disease outcomes [16], establish trends over time [17] and model the best use of healthcare resources [18, 19]. Importantly, many EHRs also provide high-quality data on medication prescribing. In claims databases, any medication claimed for under a health insurance policy is typically recorded by the insurance provider. In primary care databases, information on medications prescribed by the general practitioner (GP), such as number of tablets and dosage, are recorded, while in pharmacy databases, data on dispensing of medications are also available. Traditionally, data from EHRs have been used to assess adverse effects of treatment, especially unexpected effects. Improvements in the availability and quality of data and advances in study designs and analytical methods have broadened the value of such studies. This enables researchers to answer questions of both regulatory and epidemiological importance more quickly than with traditional study designs where data are collected in real time after conception of the study. EHRs have already been used to answer a range of questions concerning diabetes risk and treatment effects [20, 21].

Although no one database is likely to have an entire, complete picture of an individual's medical history, linkage between EHRs can improve completeness and validity of key morbidity data, as demonstrated for myocardial infarction [22], and enable the study of exposures and outcomes which would otherwise be impossible in unlinked data. In the UK, primary care data are routinely linked to Office for National Statistics death certificate data (providing detailed information on causes of death), hospital data (providing information on diagnoses from secondary care), deprivation data and disease-specific registries (e.g. for cancer, acute coronary syndromes) [12]. Similar linkages are also available between databases in the USA [23]. The availability of linked data depends greatly on the data provider, data infrastructure and, in the USA, healthcare provider. In Denmark and other Scandinavian countries, however, information across a wide range of databases (such as hospital records [11], prescriptions [24] and disease registries [25]) are all linked by a unique identity code assigned to each resident either at birth or when they become a resident [6], resulting in virtually complete population coverage and linkage. Linkages to biobanks can also provide highly detailed information on laboratory results and genetic markers (see for example http://www.bbmri-eric.eu/ (accessed 5 Jun 2017); [26, 27]. Further, although different EHRs may use differing classifications and coding systems (e.g. Read codes vs ICD), combining data from multiple sources is still possible since mappings between coding and classification systems are generally available, or may be done on a study by study basis.

Advantages of research using EHRs

- Studies are cost effective to conduct as data are already collected for other purposes
- Data are not affected by recall bias as they are collected prospectively in real time
- Data are available in near-real time, vital for a fast-changing field such as diabetes
- Large sample sizes allow for increased power to conduct granular comparisons between population subgroups and to investigate rare outcomes [7, 8]
- High validity of coded data for many diagnoses [9–11]
- Detailed prescribing and dispensing information often available for medications
- Potential for linkage across a range of healthcare settings
- Samples often representative of the source population, allowing for accurate generalisations [6, 12]

Possible pitfalls of EHRs

We summarise a broad range of issues relevant to the study of diabetes using EHRs. A previous systematic review has detailed the methodological challenges of studying glucose-lowering medications in observational studies [28]. Therefore, issues specific to the study of drug effects, such as confounding by indication (whereby the reason for being prescribed [or not prescribed] the drug is itself related to the risk of the outcome), are not covered here.

Accurate identification of diabetes status

Accurate disease ascertainment and categorisation is an essential first step towards identifying patterns of disease, and

t1.1 **Table 1** Examples of EHRs

EHR	Data types available	Examples
Primary care databases	Diagnoses of chronic and acute conditions, prescription data, information on processes of care procedures and monitoring (e.g. blood tests, BP, screening and annual health checks), as well as demographic and lifestyle information such as age, sex, smoking and alcohol consumption	Clinical Practice Research Datalink (UK) QRESEARCH (UK) SAIL database (Wales) Primary Care Sentinel Surveillance Network (Canada) Integrated Primary Care Information Database (Netherlands) The Information System for the Development of Research in Primary Care (Spain)
Secondary care databases	Admissions to inpatient, outpatient and emergency services, diagnostic and procedural codes and administrative information such as length of stay, ward and specialty area	Hospital Episode Statistics (UK) National Registry of Patients (Denmark)
Disease registries	Detailed information on the relevant condition (e.g. cancer registries have details of date of diagnosis, cancer type, grade and treatments received but may lack information on comorbidities and concomitant medication)	Primary Care Cardiovascular Database (Sweden) Global Rare Diseases Patient Registry Data Repository (USA) Myocardial Ischaemia National Audit Project (UK) Danish Huntington Register (Denmark)
Insurance claims databases	Demographic information on the individual enrolled in the insurance plan, as well as details of medical history that have been covered and medication that has been claimed for under the insurance plan (e.g. information on prescription drugs and hospital inpatient and outpatient care)	Medicare (US) Health Maintenance Organizations (HMOs) such as Molina Healthcare, Kaiser Permanente, United Healthcare (USA) National Health Insurance Research Database (Taiwan) PHARM (Italy)
Pharmacy databases	Drug dispensing, effectiveness, safety and cost data	Scottish National Prescribing System (Scotland) PHARMO database (Netherlands) Deutsches Arzneiprüfungsinstitut (Germany)
Regulatory databases	Spontaneous reports of adverse drug reactions (ADRs)	Vigibase (WHO spontaneous reports database) EudraVigilance (Europe) GECEM (France)

targeting interventions and resources appropriately. Challenges for diabetes researchers include the long latency between disease onset and diagnosis, and misclassification of diabetes type (e.g. older-onset type 1 diabetes being misclassified as type 2). Such misclassification may result in a biased estimation of the impact of diabetes on outcomes. Medication records may be used to supplement clinical data in identifying individuals with diabetes but this can present additional problems (e.g. metformin is used for the treatment of polycystic ovary syndrome and insulin is used in both type 1 and type 2 diabetes). Algorithms combining both diagnostic and supporting information (e.g. medication, laboratory results, age, BMI) have been developed to overcome these challenges [14, 29].

Differentiating between prevalent vs incident disease and treatment

In many EHRs, individuals often join the database at time points with no clear clinical significance. For example, in primary care records, the first database entry is made on the date of an individual's initial registration with the GP. At the initial visit, a GP may enter details for all pre-existing conditions. Therefore, in the period immediately after an individual enters the database, it may be unclear whether a new diabetes diagnostic code reflects existing diabetes or a new diagnosis [30]. This may limit the ability to adjust for diabetes duration, which may be an important source of confounding, particularly in studies comparing diabetes treatments. It is also typically unclear whether a new medication record in this early period reflects continuation of an existing therapy or incident use. Including prevalent users in a study of drug effects can lead to serious bias if treatment effects or risks vary over time, as is often (although not always) the case in diabetes. This is because prevalent users will have already 'survived' the early period of therapy [31]. For this reason, so-called new-user designs are generally encouraged, wherein new drug users are typically identified by requiring a certain period (e.g. 12 months) of follow-up before the first prescription [32]. However, it should be acknowledged that such designs may come at the price of loss in power, since we often reduce the sample to individuals with shorter exposure or duration of disease, which may reduce the number of long-term outcomes observed.

Use of future information

When an EHR study is designed, it is often the case that all, or a large proportion, of the follow-up information is already

available. Using future information when defining cohort inclusion, exposure status or covariate values at the time of study entry risks biasing the results because patient outcomes have influenced how they are dealt with in the study prior to their outcome [33]. As a simple example, consider a study of BMI and future risk of cardiovascular risk using a diabetes registry. Each individual may have multiple measures of BMI from the time they enter the registry until the time they exit the database or develop cardiovascular disease (CVD). If all BMI measures are used to determine whether an individual is overweight at study entry (e.g. by calculating an average BMI over follow-up), then the target comparison of 'overweight' vs 'normal weight' becomes a comparison of 'average overweight' vs 'average normal weight', leading to unclear interpretation and potential selection bias. An average normal weight could mask weight loss as a consequence of undiagnosed CVD, or a CVD diagnosis that appears late in the course of disease. Another problem of using future information is that concerning 'immortal time bias'. This term is associated with the concept that during certain time periods during follow-up, a specific outcome cannot occur. Levesque et al [34] demonstrated this using data from a Canadian health database: they defined statin users as those with 12 or more months of continuous use during follow-up, and compared rates of insulin initiation (a proxy for diabetes progression) from study entry between users and non-users. This led to an estimated protective effect of statins. The problem with this approach is that anyone experiencing the outcome (insulin initiation) before completing 12 months of statin use would be classified as a non-user as their time at risk in the study would end at this point so they could not fulfil the definition of being a statin user. The corollary to this is that those categorised as statin users could not by definition have experienced the outcome (insulin initiation) prior to starting a statin and completing 12 months of statin use, creating a period of 'immortal time' for statin users. When this event-free person-time is included in the denominator, outcome rates in the exposed group are biased downwards, leading to an overall bias towards a protective effect of exposure. When the authors instead used a correct time-updated approach wherein an individual's exposure status was updated from non-user to user once that individual reached 1 year from their first statin prescription, the protective effect of statins disappeared. Another solution might have been to start follow-up 1 year after the first statin prescription for statin users and to use a matched date for non-statin-users. Immortal time bias, along with other time-related biases, has been previously described in reference to studies of metformin and cancer risk in patients with diabetes [35] and in the previously referenced review by Patorno et al [28]. When defining inclusion criteria and exposures/covariates intended to reflect the point of study entry, it is worth asking the question 'Have I only used information that I would have had at the time of recruitment had I conducted this study in real time?' If the answer is no, then bias may inadvertently be introduced.

Dealing with the complexities of diabetes progression

One of the most common scenarios in which bias from use of future information manifests in diabetes epidemiology is when dealing with treatment switches over the course of follow-up. Studies may restrict the study population to individuals who remain on a single therapy regime throughout follow-up, leading to selection bias or immortal time. One solution is to model the treatment of interest as time-varying, thus allowing the inclusion of all patients by accounting for their treatment modality. Such a solution would be relevant to the study of any exposure (e.g. BMI, HbA_{1c}, eGFR) that changes as the disease progresses. Although an important advantage of EHRs is the ability to collect longitudinal data to investigate such time-varying exposures, dealing with confounding invariably becomes more complex in this situation. When considering how to model changes in exposure status through time, one must determine first whether information on time-varying confounders (confounders of the association between exposure and outcome that also change through time) is available in the database and second whether the time-varying confounders may also be affected by prior exposure status. If time-varying confounding is thought to be present, then adjustment for the value of the confounder at study entry only may not remove confounding for those whose exposure status changes over the course of follow-up. This can be overcome by using methods such as time-varying Cox proportional hazards models, which time-update the value of the confounder as it changes. However, if prior exposure is expected to affect future values of the confounder, then this method may not be appropriate as the adjustment may remove the effect of treatment that acts via future values of the confounder. These limitations of standard analysis methods in the presence of time-dependent confounders affected by prior exposures for diabetes research have been described in more detail in a systematic review [36], and more generally elsewhere [37]. Such issues occur both when examining time-varying treatment and time-varying risk factors such as BMI or glucose control or progressive conditions such as chronic kidney disease (CKD). For example, if we wish to examine the effect of CKD stage on mortality in individuals with diabetes, then HbA_{1c} may be a time-varying confounder of the association but CKD stage may also influence future HbA_{1c}. Methodological approaches to dealing with time-varying confounders affected by prior treatment include inverse probability weighting of marginal structural models, g-computation and g-estimation [38]. In theory, these methods correctly adjust for the time-varying confounding without losing any effect of exposure that acts via future values of the confounder, subject to certain assumptions [38]. If such methodologies are not feasible, simpler study designs in which exposures are assumed to remain fixed from study entry (analogous to intention to treat analyses) may still be used to examine exposure/outcome associations but

such designs can only answer more limited questions that ignore the reality of individuals changing treatments over time.

Finally, another consideration when dealing with time-varying exposure, is the extent to which changes in exposure are a result of reverse causality. For instance, many people lose weight shortly before diagnosis of diabetes, due to underlying ill health. Using weight measures shortly before diagnosis may lead to the erroneous conclusion that low weight is a risk factor for diabetes. It is advisable to conduct a sensitivity analysis to determine whether this may be an issue (e.g. by defining the date of exposure as being 6–12 months after the date observed within the EHR) [30].

Context in which data are collected

Understanding the purpose for which the data were initially collected and methods of data collection are critical to accurate analysis and interpretation of EHR research and for assessing the likelihood of encountering problems of missing data and unmeasured confounding.

Selection biases arising from data availability Primary and secondary care data are collected as and when individuals visit their GP or hospital and therefore samples from these databases may over-represent less-healthy individuals. This may present less of a problem in studies restricted to individuals with diabetes, since they will likely visit the GP on a semi-regular basis and thus have similar amounts and types of data recorded. However, if a general population comparison group is selected, those with available data may not be representative of the broader population. Even among individuals who do visit their GP regularly, there may be less data collected on those who are perceived to be healthier or at lower risk, as GPs are less likely to perform routine investigations in this group. Different considerations apply for claims databases: these may have an over-representation of healthier individuals, as those with pre-existing conditions may find it harder to receive medical cover.

Missing data EHR data, for the reasons outlined above, likely suffer from missing data issues. Often, we classify variables based on the presence or absence of codes. For example, when determining whether an individual has had a previous CVD event, the presence of a code will indicate 'yes', while the absence of a code will likely indicate 'no', and thus we can derive a CVD status for 100% of individuals (albeit with the possibility of misclassification). However, for measures such as blood pressure or HbA_{1c}, missing data are likely to indicate that the value has not been recorded. Analysing only the subset of individuals that have complete data on all necessary covariates is a commonly used approach but whether or not this is reasonable depends on how the missingness is associated the outcome of interest [39]. Advanced methods such as multiple imputation may be used to assess the extent to which missing data may affect the analysis and to obtain more valid estimates of association if data are missing at random, meaning that the reason for missingness is independent of the value after conditioning on other measured covariates [40]. Unfortunately, this is an untestable assumption [40, 41] and often unlikely to hold. For example, smoking is more likely to be recorded in routine primary care among smokers, and BMI is more likely to be recorded among overweight individuals. Therefore, sensitivity analysis is always advisable and there exist comprehensive practical guides to approaching analysis with missing data [42, 43]. Even if observed, data on behaviours such as smoking and alcohol consumption are unlikely to be recorded with perfect accuracy, particularly since they are often self-reported and are subject to social desirability bias [44].

Unmeasured confounding EHRs rarely contain information on diet and physical activity, which may be important confounders when looking at diabetes-related exposures and outcomes. Linkage to other sources may overcome this issue in some situations (e.g. some biobanks collect cross-sectional information on dietary intake). In some cases, the proxies may allow some degree of adjustment for unobserved variables. For example, statin use may be a reasonable proxy for high cholesterol where actual cholesterol values are not recorded. If such options are not available, a negative control can be an informative way of investigating the impact of unmeasured confounding [45]. This involves examining an association that could plausibly be affected by the same unmeasured confounders as the primary association of interest, but where the true association is expected to be null. If the result obtained is close to the known association, this provides reassurance that unmeasured confounding is unlikely to be substantially biasing the results of the primary analysis. Such a method has been successfully employed by Jackson et al in debates over influenza vaccinations [46]. The authors estimated a protective association between vaccine use and trauma hospitalisation, suggesting that unmeasured confounding may be responsible for the observed reduction in respiratory hospitalisation.

Recommendations

Although the challenges discussed in this paper were not identified systematically and were not intended to form an exhaustive list, they lead us to outline some key recommendations for best practice when studying diabetes using EHRs

Key recommendations

1. To address any question in diabetes epidemiology, we must be able to confidently identify a population of individuals with diabetes within the EHR. Consider whether algorithms combining diagnostic, therapeutic and demographic information may improve ascertainment of diabetes status, type and duration compared with the use of coded diagnostic data alone
2. Where possible, include only incident users of medications when examining treatment effects and only compare treatments that would be used at similar stages of the disease. Beyond the estimation of treatment effects, it is still important to consider whether combining prevalent and incident cases of diabetes within a study is appropriate for the question of interest
3. At any given point in time, avoid using future information to either define inclusion into the study population or to define any variable for an individual
4. Be aware of the possibility of problematic time-dependent confounding if studying a time-varying exposure (be it a treatment or otherwise) and that advanced causal methods for handling such problems tend to make strong assumptions
5. Always consider the context in which data are collected and coded when interpreting and generalising results

Although the challenges discussed in this paper were not identified systematically and were not intended to form an exhaustive list, they lead us to outline some key recommendations for best practice when studying diabetes using EHRs.

Conclusions

EHRs offer great potential for the study of complex questions beyond the scope of traditional clinical and observational studies due to the breadth and timeliness of available data and the ability for linkage to secondary care, mortality data and disease registries. As such, there is a great opportunity to allow for more accurate characterisation of diabetes type, progression of disease and quality of care.

The increasing quantity and quality of computerised health-related data offers exciting opportunities for research in diabetes. However, the danger of poor quality research with misleading results is high and could result in deleterious effects on patient care and on prescribing. Improvements in reporting of research, driven by initiatives such as the Reporting of Studies Conducted using Observational Routinely Collected Health Data (RECORD) reporting guidelines statement, may make it easier to identify the most rigorous and reliable research [47]. Further, sharing of code lists and statistical code may improve reproducibility of research using EHRs. Alongside these improvements in transparent reporting, increasing awareness of the methodological challenges, such as those outlined in this paper, is needed to help ensure that studies based on EHR data produce valid results that usefully add to the evidence base.

Funding RF and NC are funded by a Diabetes UK/British Heart foundation award (no. 15/0005250). RM is supported by a Sir Henry Wellcome Postdoctoral Fellowship from the Wellcome Trust (WT/201375/Z/16/Z). SVE is supported by a Sir George Alberti Training Fellowship (17/0005588). KB holds a Sir Henry Dale fellowship jointly funded by the Wellcome Trust and the Royal Society (107731/Z/15/Z). LS is supported by a Wellcome Trust Senior Research Fellowship in Clinical Science (098504/Z/12/Z).

Duality of interest The authors declare that there is no duality of interest associated with this manuscript.

Contribution statement All authors were involved in drafting the article and revising it critically for important intellectual content. All authors approved the final version to be published.

References

1. Edwards L, Rooshenas L, Isaacs T (2016) Inclusion of ethnic minorities in telehealth trials for type 2 diabetes: protocol for a systematic review examining prevalence and language issues. JMIR Res Protoc 5:e43

2. Hussain-Gambles M, Atkin K, Leese B (2004) Why ethnic minority groups are under-represented in clinical trials: a review of the literature. Health Soc Care Community 12:382–388
3. Zhang T, Tsang W, Wijeysundera HC, Ko DT (2013) Reporting and representation of ethnic minorities in cardiovascular trials: a systematic review. Am Heart J 166:52–57
4. Coloma PM, Schuemie MJ, Trifirò G et al (2011) Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf 20:1–11
5. Chamberlain JJ, Herman WH, Leal S et al (2017) Pharmacologic therapy for type 2 diabetes: synopsis of the 2017 American Diabetes Association standards of medical care in diabetes. Ann Intern Med 166:572–578
6. Schmidt M, Pedersen L, Sorensen HT (2014) The Danish Civil Registration System as a tool in epidemiology. Eur J Epidemiol 29:541–549
7. Brauer R, Douglas I, Garcia Rodriguez LA et al (2016) Risk of acute liver injury associated with use of antibiotics. Comparative cohort and nested case-control studies using two primary care databases in Europe. Pharmacoepidemiol Drug Saf 25(Suppl 1):29–38
8. Bhaskaran K, Douglas I, Forbes H, dos Santos-Silva I, Leon DA, Smeeth L (2014) Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. Lancet 384:755–765
9. Herrett E, Thomas SL, Schoonen M, Smeeth L, Hall AJ (2010) Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol 69:4–14
10. Wilchesky M, Tamblyn RM, Huang A (2004) Validation of diagnostic codes within medical services claims. J Clin Epidemiol 57:131–141
11. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT (2015) The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol 7:449–490
12. Herrett E, Gallagher AM, Bhaskaran K et al (2015) Data resource profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol 44:827–836
13. Shah AD, Langenberg C, Rapsomaniki E et al (2015) Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1·9 million people. Lancet Diabetes Endocrinol 3:105–113
14. Mathur R, Bhaskaran K, Edwards E et al (2017) Population trends in the 10-year incidence and prevalence of diabetic retinopathy in the UK: a cohort study in the Clinical Practice Research Datalink 2004–2014. BMJ Open 7:e014444
15. Holden SH, Barnett AH, Peters JR et al (2013) The incidence of type 2 diabetes in the United Kingdom from 1991 to 2010. Diabetes Obes Metab 15:844–852
16. Poppe KK, Doughty RN, Wells S et al (2017) Developing and validating a cardiovascular risk score for patients in the community with prior cardiovascular disease. Heart 103:891–892
17. Schmidt M, Jacobsen JB, Lash TL, Bøtker HE, Sørensen HT (2012) 25 year trends in first time hospitalisation for acute myocardial infarction, subsequent short and long term mortality, and the prognostic impact of sex and comorbidity: a Danish nationwide cohort study. BMJ 344:e356
18. Hong JL, McNeill AM, He J, Chen Y, Brodovicz KG (2016) Identification of impaired fasting glucose, healthcare utilization and progression to diabetes in the UK using the Clinical Practice Research Datalink (CPRD). Pharmacoepidemiol Drug Saf 25:1375–1386
19. Sancho-Mestre C, Vivas-Consuelo D, Alvis-Estrada L, Romero M, Usó-Talamantes R, Caballer-Tarazona V (2016) Pharmaceutical cost and multimorbidity with type 2 diabetes mellitus using electronic health record data. BMC Health Serv Res 16:394
20. Solomon DH, Massarotti GR, Lium J, Canning C, Schneeweiss S (2011) Association between disease-modifying antirheumatic drugs and diabetes risk in patients with rheumatoid arthritis and psoriasis. JAMA 305:2525–2531
21. van Staa TP, Patel D, Gallagher AM, de Bruin ML (2012) Glucose-lowering agents and the patterns of risk for cancer: a study with the General Practice Research Database and secondary care data. Diabetologia 55:654–665
22. Herrett E, Shah AD, Boggon R et al (2013) Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ 346:f2350
23. Bradley CJ, Penberthy L, Devers KJ, Holden DJ (2010) Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future. Health Serv Res 45:1468–1488
24. Kildemoes HW, Sørensen HT, Hallas J (2011) The Danish National Prescription Registry. Scand J Public Health 39(7 Suppl):38–41
25. Green A, Sortsø C, Jensen PB, Emneus M (2015) Validation of the Danish National Diabetes Register. Clin Epidemiol 7:5–15
26. Christensen H, Nielsen JS, Sørensen KM, Melbye M, Brandslund I (2012) New national Biobank of The Danish Center for Strategic Research on Type 2 Diabetes (DD2). Clin Epidemiol 4:37–42
27. Sudlow C, Gallacher J, Allen N et al (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12:e1001779
28. Patorno E, Patrick AR, Garry EM et al (2014) Observational studies of the association between glucose-lowering medications and cardiovascular outcomes: addressing methodological limitations. Diabetologia 57:2237–2250
29. Eastwood SV (2016) Algorithms for the capture and adjudication of prevalent and incident diabetes in UK biobank. PLoS One 11:e0162388
30. Lewis JD, Bilker WB, Weinstein RB, Strom BL (2005) The relationship between time since registration and measured incidence rates in the General Practice Research Database. Pharmacoepidemiol Drug Saf 14:443–451
31. Prentice RL, Langer R, Stefanick ML et al (2005) Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. Am J Epidemiol 162:404–414
32. Ray WA (2003) Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 158:915–920
33. Pocock SJ, Smeeth L (2009) Insulin glargine and malignancy: an unwarranted alarm. Lancet 374:511–513
34. Levesque LE, Hanley JA, Kezouth A, Suissa S (2010) Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ 340:b5087
35. Suissa S, Azoulay L (2012) Metformin and the risk of cancer: time-related biases in observational studies. Diabetes Care 35:2665–2673
36. Farmer RE, Ford D, Forbes HJ et al (2017) Metformin and cancer in type 2 diabetes: a systematic review and comprehensive bias evaluation. Int J Epidemiol 46:745
37. Robins JM, Hernán MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11:550–560
38. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JA (2013) Methods for dealing with time-dependent confounding. Stat Med 32:1584–1618
39. White IR, Carlin JB (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med 29:2920–2931

40. Bhaskaran K, Smeeth L (2014) What is the difference between missing completely at random and missing at random? Int J Epidemiol 43:1336–1339
41. Carpenter J, Kenward M (2012) Multiple imputation and its application. Wiley, Chichester
42. Carpenter JR, Kenward MG, White IR (2007) Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Methods Med Res 16:259–275
43. Sterne JAC, White IR, Carlin JB et al (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338:b2393
44. Kypri K, Wilson A, Attia J, Sheeran P, Miller P, McCambridge J (2016) Social desirability bias in the reporting of alcohol consumption: a randomized trial. J Stud Alcohol Drugs 77:526–531
45. Lipsitch M, Tchetgen Tchetgen E, Cohen T (2010) Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 21:383–388
46. Jackson LA, Jackson ML, Nelson JC, Neuzil KM, Weiss NS (2006) Evidence of bias in estimates of influenza vaccine effectiveness in seniors. Int J Epidemiol 35:337–344
47. Benchimol EI, Smeeth L, Guttmann A et al (2015) The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med 12:e1001885

AUTHOR QUERY

AUTHOR PLEASE ANSWER QUERY.

No Query.