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.
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

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

Ruth Farmer¹ · Rohini Mathur¹ · Krishnan Bhaskaran¹ · Sophie Eastwood² · Nishi Chaturvedi² · Liam Smeeth¹

Received: 23 June 2017 / Accepted: 24 October 2017
© The Author(s) 2017. This article is an open access publication

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.
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...
Therefore, in the period immediately after an individual enters visit, a GP may enter details for all pre-existing conditions. For example, in primary care records, the first database entry is made on the date of 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.

### 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.
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, HbA1c, 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 HbA1c may be a time-varying confounder of the association but CKD stage may also influence future HbA1c. 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 HbA1c, 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 with 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.
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.

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.

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

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
No Query.