Analyzing electronic health records: The benefits of target trial emulation

Lytske J. Bakker a,b,c,*, Lucas M.A. Goossens a,b,c, Maurice J. O’Kane d, Carin A. Uyl-de Groot a,b,c, William K. Redekop a,b,c

a Erasmus School of Health Policy and Management, ESHPM, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands
b Institute for Medical Technology Assessment, IMTA, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands
c Erasmus Centre for Health Economics Rotterdam (EsCHER), Erasmus University, Rotterdam, the Netherlands
d Clinical Chemistry Laboratory, Altnagelvin Hospital, Western Health and Social Care Trust, Limavady Rd, Londonderry BT47 6LS, Londonderry, N. Ireland, UK

A R T I C L E  I N F O
Keywords:
Electronic health records
Observational research
Type 2 diabetes
Information technology

A B S T R A C T

Aim: Electronic health records (EHRs) are increasingly used in effectiveness and safety research. However, these studies are often at risk of bias. This study demonstrates the relevance, and discusses challenges, of using target trial emulation to avoid bias, such as selection bias, immortal time bias and confounding when performing observational research with EHRs.

Methods: Target trial emulation can be used to identify and address some of the drawbacks of observational research in a systematic way. Potential sources of bias are identified by describing key components of an ideal randomized controlled trial and comparing this to the observational study actually performed. The methods were applied to assess treatment response to antidiabetic treatment using EHRs from patients with diabetes treated in secondary care.

Results: Using target trial emulation ensured prevalent users were excluded and patients were not included based on information generally not available when initiating a clinical trial. Furthermore, applying these methods demonstrated how the number of records eligible for use can rapidly decrease. Hereafter, adjustments were performed to address potential sources of bias and it was shown that missing variables essential for adjustment can be an important issue.

Conclusions: Using target trial emulation, sources of selection bias and confounding were identified and adjusted for accordingly when analysing treatment response in patients with type 2 diabetes. However, when using EHR data to emulate a target trial, samples containing sufficient information on outcome measures and variables to adjust for confounding and selection bias are essential given the risk of missing data.

Public Interest Summary

Results from research that uses electronic health records (EHRs) are always at risk of being biased due to selection bias, immortal time bias and confounding. Target trial emulation is a systematic way of identifying and avoiding different types of bias in these studies. We demonstrate the value, but also important challenges, of using target trial emulation when performing research using data from EHRs. The methods were applied to assess treatment response to antidiabetic treatment using EHRs from patients with diabetes treated in secondary care. Using target trial emulation, several sources of bias were identified and adjusted for accordingly when analysing treatment response in patients with type 2 diabetes. However, when using EHR data to emulate a target trial, samples containing sufficient information on outcome measures and variables to adjust for bias such as confounding and selection bias are essential given the risk of missing data.

Introduction

The randomized controlled trial (RCT) is the preferred method for assessing the efficacy of novel treatments. Randomly assigning patients to either the treatment group or the comparator enables researchers to isolate the effect the treatment has on the outcome. However, RCTs are often costly to perform, have a limited generalizability and may pose ethical challenges [1,2]. Furthermore for diseases such as diabetes, where many different treatment combinations are possible, performing an RCT for each possible antidiabetic treatment combination is often not feasible. The use of observational research and electronic health records (EHR) to assess the real-world effectiveness of these treatments has sometimes been suggested as an alternative [1]. However, observational studies can be challenging to perform due to missing data and the risk of bias [1,3–7]. Even though best practice methods have been emphasized

* Corresponding author: ESHPM, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands.
E-mail address: bakker@eshpm.eur.nl (L.J. Bakker).

https://doi.org/10.1016/j.hlpt.2021.100545

Available online 6 July 2021
2211-8837/© 2021 Published by Elsevier Ltd on behalf of Fellowship of Postgraduate Medicine.
for observational research in a chronic disease such as diabetes [8], recent systematic reviews have found that many observational studies are still at risk of bias such as selection bias, immortal time bias and confounding [1,5,7].

Target trial emulation can be used to identify and address some of the drawbacks of observational research in a systematic way [2]. In target trial emulation, potential sources of bias are identified by describing an ideal trial and comparing this to the observational study that is designed to emulate this target trial [9]. If shortcomings of the observational study that have a large impact on the quality of the study cannot be overcome, researchers can adjust their design or find additional data [9]. Elements of target trial emulation (i.e. eligibility criteria) are sometimes already presented and used in observational research [2,8,9]. However, these elements are often not used in a systematic way and the necessity of using them, and the challenges that can be expected, may not always be apparent to researchers [2]. In this paper, we aim to offer practical guidance for those performing real-world effectiveness and safety research using EHRs. We demonstrate the value of target trial emulation but also discuss several challenges that can be expected by emulating a target trial using EHRs of patients with diabetes.

**Methods**

**Identifying bias**

Throughout the remainder of the paper we will discuss the components of target trial emulation based on Hernán et al. [2] and demonstrate how these can be used to systematically reduce the risk of bias. There are many ways in which results from observational research are at risk of bias. While the definitions of these types of bias may differ between research disciplines, they can be clearly illustrated using directed acyclic graphs (DAGs). A DAG presents all assumed potential causal relationships between variables using directional arrows [10]. Thus, when an arrow is drawn from the exposure variable (i.e. treatment) to the outcome variable (i.e. Hemoglobin A1C (HbA1c) response) the arrow cannot also go the other way (Fig. 1a). Moreover, DAGs are acyclic, thus no variable can cause itself. DAGs may contain both measured and unmeasured causal variables as well as common causes and effects of exposure and outcome [11]. If the rules for drawing DAGs are followed, the graphs can be used to determine which statistical adjustments are required. Statistically, an exposure and outcome are associated when the DAG contains an open connection (‘path’) between them [10]. These paths can be closed or opened by applying statistical adjustments, conditioning or altering the research design.

**Target trial emulation**

In target trial emulation, the seven components of the target trial protocol are compared to the observational study to be performed. The first component of the protocol is to define eligibility criteria using only data that would be available prior to treatment initiation [2]. In the target trial, information on follow-up would not be available prior to starting treatment and using this information for patient selection could induce selection bias. Selection bias can occur when conditioning on a common effect of outcome and exposure (Fig. 1c & 1d) [11]. This opens a path between exposure and outcome, which had been closed due to the presence of the common effect (‘collider’). Studies are at risk of selection bias when future information is used to define inclusion and exclusion criteria. In research that uses EHRs, a frequently seen example of this is the selection of patients based on the availability of outcome data (Fig. 1c) [1,6,12]. It is possible that there is no difference in HbA1c reduction between treatment with an sodium glucose transporter-2 inhibitors (SGLTs) vs a dipeptidyl-peptidase 4 inhibitors (DPP4s) but patients receiving SGLTs are at risk of more severe side effects of treatment, which may reduce attendance at follow-up visits [13]. Meanwhile, depression may independently result in loss to follow-up and higher HbA1c values [14,15]. Missing patient data in a patient prescribed an SGLT may therefore relate both to side effects of treatment and to depression. Thus, when including only those with data at follow-up we would open the backdoor path from treatment to HbA1c at follow-up through depression. We would find that patients included in follow-up receiving DPP4s have lower HbA1c values while actually there is no difference in HbA1c.

The second component of target trial emulation describes the treatment strategies, preferably including only new-users [2]. By including all events that occur early after drug initiation, a new-user design reduces the risk of selection bias [16]. This form of selection bias, also

![Fig. 1. Directed acyclic graphs that represent the causal relationships between treatment and treatment response (HbA1c at follow-up). A) Assumes that the effect of treatment on HbA1c response is only decided by treatment. B) Confounding: Assumes that the selected treatment and HbA1c response share a common cause, namely baseline weight. C) Selection bias: Patients are included when follow-up data is present. If in truth no relation between treatment and HbA1c exists, a spurious relation would be induced by selecting on a collider thus opening the backdoor path through depression. D) Prevalent user bias: Here we assume no difference exists in HbA1c at follow-up between dipeptidyl-peptidase 4 inhibitors (DPP4s) and sodium glucose transporter-2 inhibitors (SGLTs). However, if DPP4s are more often prescribed prior to hospital referral and only those users with a good response are still on DPP4s when referred to the hospital, this would open the backdoor path through the initial HbA1c measurement ‘HbA1c T1’ by selecting on the collider ‘Prescribed in hospital’. E) Immortal time bias: In truth no relation exists between treatment and time to progression defined as elevated HbA1c measurements. If inclusion criteria require patients on an SGLT receive treatment for at least 1 month but follow-up starts at treatment assignment, results would be biased upwards for SGLTs. HbA1c T1 = HbA1c at time 1, HbA1c T2 = HbA1c at time 2.](image-url)
referred to as prevalent user or survival bias, occurs when patients are included in the study that were already prescribed the drug prior to the start of follow-up (Fig. 1d). Thus, if we were to include patients already prescribed treatment by the general practitioner prior to hospital referral then some patients that already failed treatment early after initiating therapy would be excluded. If DPP4s are more often prescribed by the general practitioner than SGLTs, this could lead to an underestimation of the benefits of SGLTs over DPP4s.

The third component contains the assignment procedures to reduce the risk of confounding. Where random assignment is often the preferred strategy in the target trial, adjustment for potential (post-) baseline confounders is required in observational research [2]. Confounding is present when the exposure and outcome share a common cause (Fig. 1b) [10]. In Fig. 1b, it is assumed that treatment and HbA1c response are both (partially) influenced by a patient’s weight, measured before treatment. This opens a path between treatment and outcome that does not represent the causal effect of treatment. Failing to adjust for weight in such an analysis could result in incorrect conclusions that treatment improves HbA1c while in truth weight causes both the exposure and the outcome.

The next component contains the follow-up period defined by the start and the end of follow-up. When initiation of follow-up is not aligned with eligibility criteria and treatment assignment, immortal time bias can influence results [17]. Immortal time is the time in which the outcome cannot occur [6,18]. Suppose we wish to compare time to progression according to elevated HbA1c in patients prescribed an SGLT compared to DPP4s. If inclusion criteria requires patients on SGLTs receive treatment for at least 1 month because of the intensification schedule, but follow-up starts when the drug is first prescribed, then results are at risk of immortal time bias. The time to progression will be higher for patients in the SGLT arm since they receive additional time in which they cannot progress at the beginning of follow-up (Fig. 1e). Excluding prevalent users and avoiding use of future information can ensure alignment of inclusion criteria, treatment assignment and follow-up, aiming to avoid immortal time bias [6].

This is followed by a description of the relevant outcome for which blinded measurements are often preferred [2], but often not possible in the observational study. The causal contrast of interest is then selected in the following component, which often includes the intention-to-treat and/or the per-protocol effect [2]. An intention-to-treat analysis includes all patients in the analysis within the arm to which they were originally randomized, irrespective of whether they completed treatment. A per-protocol effect includes only those patients that completed treatment. Both contrasts may require specific statistical adjustments.

In the final component, the analysis plan reports the analyses that need to be performed to properly estimate the causal contrasts of interest. Statistical techniques such as inverse probability weighting, imputation, stratification, g-methods and regression can be used to adjust for (postbaseline) selection bias and confounding [2,10,12].

We applied target trial emulation to an example from diabetes care by assessing the effectiveness of SGLTs compared to DPP4s added to insulin therapy in patients with type 2 diabetes.

**Patient level data**

To demonstrate several benefits and challenges of target trial emulation, hospital EHRs were used from patients with diabetes treated between June 2012 and December 2017 at the Western Health and Social Care Trust in Northern Ireland. The EHR was a specific diabetes management system: Diamond (Hicom, Surrey, UK). EHR patient demographic information was populated from a patient administration system. Laboratory test results were transferred automatically from the local Laboratory Information Management System. Clinical data were entered live at the time of the patient consultation by the clinical medical or nursing staff; this included the recording of updated anthropometric data, medication changes or the development of new clinical problems. In the records, any active ingredient and/or trade name mentioned were given an individual identifier. These were automatically classified into subgroups (i.e. antidiabetic drugs, antibiotics etc.). This list was checked manually to identify any drugs wrongly classified and active antidiabetic drugs were grouped according to drug class. The initiation date of a drug combination was the first date on which the combination was recorded. The end date was the first date of the record on which the combination was no longer recorded after initiation.

**Target trial emulation applied to EHRs in diabetes care**

In Table 1, we present the target trial protocol alongside the design of the observational study.

**Eligibility criteria**

The target trial is an RCT including patients with type 2 diabetes managed in secondary care who do not achieve their personalized HbA1C target after at least 12 weeks of treatment with insulin with or without oral antidiabetics (metformin/sulfonylurea). Thus, these are patients receiving insulin plus an oral antidiabetics (metformin/sulfonylurea) for at least 84 days.

**Table 1**

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Target trial</th>
<th>Observational study using EHR data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Eligibility criteria</strong></td>
<td>- Patients with type II diabetes treated in secondary care</td>
<td>- Patients with type II diabetes treated in secondary care</td>
</tr>
<tr>
<td></td>
<td>- Patients receiving insulin ± an oral antidiabetics (metformin/sulfonylurea) for 84 days or more</td>
<td>- Patients received insulin ± an oral antidiabetics (metformin/sulfonylurea) for at least 84 days.</td>
</tr>
<tr>
<td><strong>Treatment strategies</strong></td>
<td>An SGLT compared to a DPP4 inhibitor with insulin ± an oral antidiabetics (metformin/sulfonylurea).</td>
<td>Similar to target trial</td>
</tr>
<tr>
<td><strong>Assignment procedures</strong></td>
<td>Patients are randomly assigned to an SGLT or a DPP4 in combination with insulin ± an oral antidiabetics (metformin/sulfonylurea).</td>
<td>Decided by physicians and their patients on a case-by-case basis.</td>
</tr>
<tr>
<td><strong>Follow-up period</strong></td>
<td>Starts at randomization.</td>
<td>Adjusted for several confounding variables or their proxies measured at baseline i.e. weight, HbA1c, duration of diabetes, age, insulin regimen, number of non-diabetes drugs.</td>
</tr>
<tr>
<td></td>
<td>Ends at 6 months of follow-up, when lost to follow-up or death</td>
<td>Started on the prescription date for the SGLT or DPP4. Ended at first HbA1c measurement at least 3 months after drug initiation, when lost to follow-up or death.</td>
</tr>
<tr>
<td><strong>Outcome</strong></td>
<td>Blinded measurement of HbA1c 6 months after randomization.</td>
<td>Blinded first measurement of HbA1c &gt; 3 months after treatment initiation.</td>
</tr>
</tbody>
</table>

Measurement was blinded as this was not performed by the prescribing clinician.

EHR = Electronic Health Record, SGLT = sodium glucose transporter-2 inhibitors, DPP4 = dipeptidyl-peptidase 4 inhibitors.

Fig. 2. The number of prescriptions that remain when applying relevant selection criteria to the initial 7,927 patients. The column on the right hand side illustrates the variation in the next prescribed treatment for the remaining 10% that received insulin ± an oral antidiabetic after applying selection criteria. OAD = Oral antidiabetic drug.
generally not newly diagnosed patients with diabetes but more complex, progressed patients seen by specialists in secondary care. The eligibility criteria in the observational study were identical except that we explicitly assumed that all those receiving intensification do not achieve their personalized targets and that the personalized target itself does not influence the treatment prescribed. We refrained from using the availability of Hba1c follow-up measurements as a criteria for inclusion since this would have also been unknown upon inclusion in the target trial.

These criteria were applied to the EHRs available and Fig. 2 illustrates how the number of patients eligible for inclusion can greatly reduce with each criteria. The initial dataset consisted of 183,570 prescription records of 7,927 patients. This number reduced with almost 90% after applying selection criteria. Only 57% of patients remained eligible after applying selection criteria. Only 57% of patients remained when only patients with type 2 diabetes were included, 21.3% received treatment with an insulin ± an oral antidiabetic and 12.8% received this treatment for at least 84 days. The treatment that followed varied greatly since a wide variety of diabetes treatment combinations are available and treatment decisions depend on the patient and hospital. Of course, the number of patients for whom relevant baseline and follow-up data was missing is not yet considered here but this is known issue when using EHR data.

**Treatment strategies**

In both the target trial and the observational study, eligible patients are randomly assigned to an SGLT or a DPP4 added to insulin with or without an oral antidiabetic (metformin/sulfonylurea). In this example, we did not limit the strategy to a specific dose. However, if the aim is to evaluate specific dosing schedules it should be considered that missing information on dose prescribed can be an important issue [19].

In Table 2, we present the sample of patients that fulfilled eligibility criteria and received the relevant treatment strategy. By including only patients that were on a treatment with insulin ± an oral antidiabetics for at least 84 days we did not include any prevalent users. Patients that received an SGLT were generally younger, had a shorter duration of diabetes and were heavier. HbA1c within 6 months prior to treatment initiation was similar for both groups however, the number of missing values was high. The time between measuring weight and HbA1c prior to treatment and initiating treatment was similar between the two groups. Furthermore, patients that received an SGLT were more often treated with a combination of basal and bolus insulin compared to patients that received a DPP4.

### Assignment procedures

While in the target trial, patients are randomly assigned to either treatment, in the observational study statistical adjustments, selected based on subject knowledge, would be used to emulate randomization. When drawing a DAG to determine for which variables should be adjusted, all potential confounders should be included. Fig. 3 shows an example of a DAG of the assumed causal relationship between treatment and HbA1c response. Here potential baseline confounders would be duration of diabetes, HbA1c, age, glomerular filtration rate (eGFR), frailty, insulin resistance and weight. It is evident that in the observational study missing data on confounders can be an important issue (Table 2) in addition to variables being absent altogether. Sometimes confounders are not available in the data and alternative measures to acquire the data such as linking records from different settings have not yielded sufficient input. In that case, proxies can sometimes be used to adjust for bias [20]. For instance, in the EHRs from which the sample was selected, eGFR, insulin resistance and frailty were unmeasured confounders. Here type of insulin (basal vs bolus vs basal-bolus insulin) [21]; weight and duration of diabetes [22] could be included as proxy variables for insulin resistance in subjects with type 2 diabetes. The

<table>
<thead>
<tr>
<th></th>
<th>Received DPP4 (N = 160)</th>
<th>Received SGLT (N = 123)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex, n(%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>77 (48)</td>
<td>51 (41)</td>
</tr>
<tr>
<td>Male</td>
<td>83 (52)</td>
<td>72 (59)</td>
</tr>
<tr>
<td>Age, n, mean(sd)</td>
<td>134; 70.29 ± 10.10</td>
<td>110; 57.30 ± 10.62</td>
</tr>
<tr>
<td>Duration of diabetes in years, n, mean (sd)</td>
<td>134; 17.54 (7.89)</td>
<td>110; 13.22 (6.19)</td>
</tr>
<tr>
<td>Weight in kg, n, mean(sd)</td>
<td>116; 87.64 (20.01)</td>
<td>112; 105.00 (21.69)</td>
</tr>
<tr>
<td>Days from baseline weight to treatment, n, median(IQR)</td>
<td>116; 36.50 (20-112)</td>
<td>112; 33.50 (20-56.25)</td>
</tr>
<tr>
<td>Baseline HbA1c in mmol/mol, n, mean (sd)</td>
<td>82; 78 (17)</td>
<td>73; 80 (20)</td>
</tr>
<tr>
<td>Baseline HbA1c in %, n, mean(sd)</td>
<td>82; 9.3 (1.6)</td>
<td>73; 9.5 (1.8)</td>
</tr>
<tr>
<td>Days to HbA1c measurement, n, median(IQR)</td>
<td>82; 36.50 (6.25-70.25)</td>
<td>73; 33.50 (20-56.25)</td>
</tr>
<tr>
<td>HbA1c at follow-up in mmol/mol, n, mean(sd)</td>
<td>77; 72 (16)</td>
<td>74; 74 (17)</td>
</tr>
<tr>
<td>HbA1c at follow-up in %, n, mean(sd)</td>
<td>77; 8.7 (1.5)</td>
<td>74; 8.9 (1.6)</td>
</tr>
<tr>
<td>Receive antidepressants, (%)</td>
<td>Yes 6 (4)</td>
<td>9 (7)</td>
</tr>
<tr>
<td>No 154 (96)</td>
<td>114 (93)</td>
<td></td>
</tr>
<tr>
<td>Number of non-diabetes drugs, mean (sd)</td>
<td>1.54 (1.98)</td>
<td>1.71 (1.97)</td>
</tr>
<tr>
<td>Insulin Type, n(%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basal and bolus</td>
<td>111 (69)</td>
<td>104 (85)</td>
</tr>
<tr>
<td>Basal or bolus</td>
<td>49 (31)</td>
<td>19 (15)</td>
</tr>
<tr>
<td>Sulfonylurea, n(%)</td>
<td>Yes 26 (16)</td>
<td>21 (17)</td>
</tr>
<tr>
<td>No 134 (84)</td>
<td>102 (83)</td>
<td></td>
</tr>
</tbody>
</table>

DPP4 = dipeptidyl-peptidase 4 inhibitor, SGLT = sodium glucose transporter-2 inhibitor, IQR = interquartile range, sd = standard deviation.

![Fig. 3. Directed Acyclic Graph representing the underlying causal structure of the difference in response between patients receiving a sodium glucose transporter-2 inhibitor or a dipeptidyl-peptidase 4 inhibitor added to insulin ± oral antidiabetics. Variables in red were not measured and when available, proxy variables were used. eGFR = glomerular filtration rate.](image-url)
number of non-diabetes drugs a patient received prior to treatment initiation could serve as a proxy for frailty [23]. For recent eGFR results, no suitable proxy would be available and thus this would remain an unmeasured confounder. eGFR is an unmeasured confounder since patients with an eGFR between 30-60 ml/min would be eligible for receiving a reduced dose of some DPP4s but not SGLTs. At this point it can then be considered whether it is worth continuing with the study, given the absence of such an important confounder.

Another important point when including baseline confounders is the time at which these variables are measured. Ideally, a baseline HbA1c measurement is available from a date as close as possible before the date of novel treatment initiation. Therefore, an important difference between the observational study and target trial would therefore be that baseline HbA1c is defined as the most recent measurement within 6 months prior to treatment initiation, and weight is defined as the most recent measurement within 365 days before treatment initiation. It would be expected that a recent HbA1c measurement is available given that an elevated HbA1c will often be a reason to switch treatments. However, in the records included in the example, routine HbA1c measurements could also be performed in primary care and therefore not directly available in secondary care EHRs. For HbA1c, only 14% of available measurements were obtained more than 3 months before initiating treatment.

Follow-up period

In the target trial, follow-up starts at randomization and ends at six months of follow-up, loss to follow-up or death. The follow-up period of the observational study is similar and starts on the date that the first prescription of the treatment combination is recorded, and ends at the first follow-up visit more than 3 months after treatment initiation, at 12-months after treatment initiation or death.

Outcome

In the target trial, researchers are blinded when measuring the outcome, HbA1c in mmol/mol at 6 months after randomization. However, in the observational study the follow-up visit is unlikely to take place at 6 months after treatment initiation and a wider time interval would be adopted, including the first follow-up visit recorded between 3-12 months after treatment initiation. Such differences in time to measurement may influence results if there are differences between the two treatment arms in time to first follow-up. For a disease such as diabetes HbA1c values generally tend to increase over time and thus if one of the arms has a longer time to first measurement it could be that these measurements are biased upwards simply because of the time to measurement. In the patient sample in Table 2, the time between treatment initiation and follow-up measurement is similar between the two arms.

Causal contrast of interest

In a target trial, often the intention-to-treat effect as well as the per-protocol effect are estimated. Here in the observational study, only the estimation of the intention-to-treat effect would be possible. In the EHR, only prescription data is available and thus there is no insight concerning whether patients received and ingested their medication.

Statistical analysis plan

When estimating the intention-to-treat effect in the observational study, adjustments would be needed for baseline confounding (Fig. 3) and selection bias due to loss to follow-up (Fig. 1c). HbA1c measurements between 3-12 months would be included as the outcome variable and the type of drug prescribed (SGLT or DPP4) would be the exposure variable. Confounders would include the following baseline variables:

- HbA1C, weight, type of insulin, number of non-diabetes drugs, age and diabetes duration and whether or not a patient receives antidepressants would be used to adjust for selection bias.

Adjusting for confounding could be achieved by for instance regression, and adjusting for selection bias due to loss to follow-up could be achieved by inverse probability weighting [24] or multiple imputation [12]. Inverse probability weighting has been recommended because it is valid and less complex to perform, whereas multiple imputation is considered complex but also efficient and has been previously recommended when analysing EHR data [1,12]. When using inverse probability weighting to adjust for loss to follow-up related to depression, weights can be based on a logistic regression model that estimates the inverse of the probability of being lost to follow-up while including whether or not a patient receives antidepressants at baseline. The 95% confidence intervals can be estimated using a bootstrap variance estimator.

The second option would be to use multiple imputation which is generally accepted as one of the better practices for handling missing data [1,12,25,26]. In multiple imputation, the distribution of the observed data is used to impute realistic values for the missing data [25]. Multiple imputation using chained equations is often used in which a model is estimated for each variable containing missing data [25]. These missing variables are then repeatedly imputed and Rubin’s rules can be used for pooling the imputed datasets [26]. Sensitivity analyses can then be performed in which imputed variables are varied for instance up to 50%.

All variables included in the analysis model should also be included in the imputation model as well as variables used to adjust for selection bias. Auxiliary variables (i.e. sex and baseline height) can be included in the model to improve model fit. Auxiliary variables are those that are not included in the final analysis but are used for imputation because they provide information on the missing variables, increasing the likelihood that the ‘missing at random’ assumption holds [25,27]. The fraction of missing information can be used to decide which auxiliary variables to include. The fraction of missing information is high when the variables in the dataset provide limited information to impute the missing values [27]. When performing multiple imputation using chained equations, predictive mean matching is often used for larger sample sizes and imputation of continuous variables [25]. However, other methods such as logistic regression and weighted predictive mean matching (MIDASTouch) can be selected for specific types of variables (i.e. nominal) or with smaller samples [26]. The number of imputations (m) can be estimated depending on the fraction of missing information and the proportion of missing data. With for instance the observational study, data are assumed to be missing at random and missingness therefore depends on covariates measured in the sample. Furthermore, in the example the sample size is small, the fraction of missing information is high and the proportion of missing data is high, therefore, the number of imputations should be high (m = 70) and a method for small samples such as MIDASTouch can be used [26,28].

After adjusting for selection bias using either multiple imputation or inverse probability weighting regression can be performed to adjust for confounding. To estimate effects for the observational study a generalized linear model with a gamma distribution and log link can be used, results are then presented as the relative change in HbA1c when receiving an SGLT compared to a DPP4 (Table 3). Thus, if the HbA1c after 6 months of follow-up for a patient is 9.2% (77 mmol/ml) when receiving a DPP4, a coefficient of -10% would imply that this same patient would have had an outcome of 8.5% (69 mmol/ml) had they received an SGLT. In the example, no evidence of a clinically meaningful difference in effect can be found, estimated as a 1.7% relative increase in HbA1c when receiving an SGLT compared to a DPP4 with considerable uncertainty (95% CI: -5.4%–9.4%) (Table 3). When adjusting for confounding, small differences can be seen after correcting for loss to follow-up using inverse probability weighting (1.0% change, 95% CI: -8.5%–11.5%) and after multiple imputation (-0.3% change, 95% CI:
Missing data can limit feasibility of research and bias (i.e. selection bias) is often not properly addressed [1, 5]. However, using these EHRs for research purposes remains challenging.

### Table 3

<table>
<thead>
<tr>
<th>Relative change in HbA1C</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adjustment</td>
<td>1.7% [-5.4%; 9.4%]</td>
</tr>
<tr>
<td>Adjustment for confounding,</td>
<td>1.4% [-7.2%; 10.8%]</td>
</tr>
<tr>
<td>Adjustment for loss to follow-up</td>
<td>1.0% [-8.5%; 11.5%]</td>
</tr>
<tr>
<td>Adjustment for confounding, using IPW</td>
<td>-0.3% [-7.5%; 7.5%]</td>
</tr>
</tbody>
</table>

IPW = Inverse probability weighting, MI = Multiple Imputation

-7.5%/-7.5%). Both report small differences between the two drugs with wide confidence intervals. The effect estimated using multiple imputation varies from -5% to 3% throughout sensitivity analyses and including the use of a sulfonylurea and smoking as auxiliary variables does not lower the fraction of missing information.

### Discussion

The use of EHRs has resulted in many opportunities for real-world effectiveness and safety research for chronic diseases such as diabetes. However, using these EHRs for research purposes remains challenging. Missing data can limit feasibility of research and bias (i.e. selection bias and confounding) is often not properly addressed [1, 5-7] which limits the value of results for clinical practice. Here, target trial emulation can be used to systematically identify potential sources of bias and adjust results accordingly.

We discussed the use of target trial emulation and apply this to an example for diabetes research. By identifying differences between the target trial and the observational study performed, important sources of bias can be identified. Frequently recurring pitfalls, such as including patients based on future information and including prevalent users, can be avoided by comparing eligibility criteria of the target trial with those of the observational study. Furthermore, by identifying differences in the assignment procedures and outcomes between the target trial and observational study, potential sources of bias can be identified and adjusted where needed. However despite its advantages, target trial emulation should be used thoughtfully. Some sources of bias, though identified, cannot be addressed (i.e. absence of eGFR data) and sometimes decisions to reduce bias and improve internal validity can limit generalizability and reduce sample size. Where strict inclusion criteria reduce the risk of bias, they also reduce generalizability. Furthermore, correcting for certain confounders in the treatment assignment step requires sufficient data to be available on confounders. Use of imputation could address issues with missing confounders but is complex to apply and could also lead to bias when used incorrectly. As with all observational studies, it is unlikely that researchers are able to exclude all potential sources of bias. Therefore, they will carefully need to consider whether results are still of clinical value despite the potential of residual bias to be present.

The amount of missing data is an important problem that influences many steps of target trial emulation using EHRs. Missing data are an unavoidable fact of life in the realm of EHRs and registries. Target trial emulation stimulates researchers to critically consider which variables are essential to generate valid inferences when drafting the protocol for the target trial. In the example, data on a relevant confounder (eGFR) was absent and it is uncertain whether continuing is worthwhile in the presence of residual confounding and the study could also be halted until additional data is collected. To improve validity of results, we recommend researchers formulate expectations concerning the proportion of missing data as well as the availability of variables needed for imputation or weighting (i.e. outcome variables, exposure variables, confounders, causes of selection bias and potential auxiliary variables) when using EHR data. When large amounts of missing data are to be expected, collecting additional data would be the preferred option. Ideally, EHRs are available from for instance initiatives such as the Clinical Practice Research Datalink in which only those primary care practices are included that adhere to certain quality standards in terms of reporting, and linking to for instance data from secondary care is possible.

Selecting variables to include in DAGs can be challenging and depends on the setting and time [29]. For instance, physicians in the US are likely to consider a patient’s social economic status (SES) in treatment selection whereas in Northern Ireland no out-of-pocket payments are required to receive SGLTs or DPP4s. Therefore, SES was excluded as a confounder in the analyses. Medication adherence was also not included but could be an important confounder in many other settings. When including variables such as medication adherence as a confounder across various treatments, the value of this confounder will depend on the time of measurement [5]. When adjusting for time-dependent confounders, appropriate statistical analyses should be used such as g-estimation [6]. It should be recognized that obtaining information on medication adherence from EHRs can be challenging [30], and it might therefore often be included as an unmeasured confounder. In our analyses, eGFR was considered an important unmeasured confounder but sometimes proxies can be used. In the example in this study, we included proxies for insulin resistance and frailty. However, when interpreting results it should be recognized that proxies are often imperfect and therefore residual bias can remain. Furthermore, proxies should be used with care since their use can also open backdoor paths, thus introducing bias [20].

However, generally some data will remain missing despite efforts of additional data collection and researchers will need to perform analyses while taking into account this missing data. Multiple imputation when analysing EHR data is often recommended since it is one of the few methods that is both efficient and effective in reducing bias, if used correctly [1, 12, 25]. However, results from multiple imputation can also be biased when the likelihood of model misspecification is large for instance when the proportion of missing data is high [31]. In the example provided in this study, we illustrated that the large amounts of missing data that can be present in EHRs might limit the value of imputation when other (auxiliary) variables have a limited predictive value and are themselves also missing. The proportion of missing data and the fraction of missing information were high while correlations with many of the variables included were low. When auxiliary variables contain many missing values, the risk of bias in regression estimates increases because of the higher proportion of missingness and ratio of variables to complete cases [32]. In such scenarios, inverse probability weighting might be preferred [31]. However, both strategies assume data are missing at random and when this assumption does not hold and data are missing not at random both methods could lead to biased estimates [31]. When collecting additional information or performing statistical adjustment such as inverse probability weighting or multiple imputation are not possible, researchers can consider whether continuation is worthwhile. For instance, in the example used in this study, the absence of information on eGFR could be a motivation to cease the study altogether. In other cases when adjustments cannot be made for all sources of bias but their impact is considered small, acknowledging their presence could assist interpretation of results.

In this study, we used EHRs of the Western Health and Social Care...
Trust to discuss several benefits and challenges when using target trial emulation. The dataset presented several challenges one of which was the small size of the sample after applying eligibility criteria. The required sample size to assess significance of the effect found in this study is 868 patients per arm assuming an alpha of 0.05 and a beta of 0.8. Thus, if the same amount of missing data is expected, EHRs of roughly ten times as many patients (n = 71,673) would be required before applying eligibility criteria. Furthermore, it should be noted that there are many more aspects that should be considered when performing an observational study using EHRs relating to for instance data extraction, data pre-processing and data validation not addressed in this study [33].

Observational research that uses EHRs to assess real-world effectiveness and safety is crucial for informing clinical practice. Target trial emulation is a useful tool for conducting these studies since it enables researchers to avoid frequently recurring problems such as selection bias, immortal time bias and confounding. However, researchers should consider that emulating a target trial using EHRs can be challenging when large variations in treatments prescribed are expected and the amount of missing data is large. Therefore, target trial emulation can only be used to improve care when sufficient information on outcome measures and variables to adjust for bias is collected.

Sources of Funding

This work was supported by European Union’s Horizon 2020 Research and Innovation Programme grant number 644906. The study sponsor/funder was not involved in the design of the study; the collection, analysis, and interpretation of data; writing the report; and did not impose any restrictions regarding the publication of the report.

Declaration of Competing Interest

Carin Uyl-de Groot reports unrestricted grants from Boehringer Ingelheim, Celgene, Janssen-Cilag, Genzyme, Astellas, Sanofi, Roche, Astra Zeneca, Amgen, Gilead, Merck, Bayer, outside the submitted work. The remaining authors have no competing interest to declare.

Ethical Approval

The analyses were performed as part of the AEGER project and ethical approval was granted by the Chelsea Research Ethics Committee (REC reference 16/LO/2018).

Acknowledgements

We would like to thank N. Philip, A. Bakhityar, R. Manzoor (all Kingston University) & J. Rumbold (Nottingham Trent University) in their assistance in extracting data from the Hicom Server.

Data Statement

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

References