Real-world evidence: the devil is in the detail

Mugdha Gokhale[1,2] · Til Stürmer[2] · John B. Buse[3]

Received: 7 January 2020 / Accepted: 1 May 2020 / Published online: 15 July 2020

Abstract

Much has been written about real-world evidence (RWE), a concept that offers an understanding of the effects of healthcare interventions using routine clinical data. The reflection of diverse real-world practices is a double-edged sword that makes RWE attractive but also opens doors to several biases that need to be minimised both in the design and analytical phases of non-experimental studies. Additionally, it is critical to ensure that researchers who conduct these studies possess adequate methodological expertise and ability to accurately implement these methods. Critical design elements to be considered should include a clearly defined research question using a causal inference framework, choice of a fit-for-purpose data source, inclusion of new users of a treatment with comparators that are as similar as possible to that group, accurately classifying person-time and deciding censoring approaches. Having taken measures to minimise bias 'by design', the next step is to implement appropriate analytical techniques (for example propensity scores) to minimise the remnant potential biases. A clear protocol should be provided at the beginning of the study and a report of the results after, including caveats to consider. We also point the readers to readings on some novel analytical methods as well as newer areas of application of RWE. While there is no one-size-fits-all solution to evaluating RWE studies, we have focused our discussion on key methods and issues commonly encountered in comparative observational cohort studies with the hope that readers are better equipped to evaluate non-experimental studies that they encounter in the future.

Keywords Active comparator · Bias · Data · Evidence · Healthcare · New-user · Pharmacoepidemiology · Real-world · Review · Study design

Abbreviations

DPP-4	Dipeptidyl peptidase-4
RWD	Real-world data
RWE	Real-world evidence
SGLT2	Sodium–glucose cotransporter-2

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00125-020-05217-1) contains a slideset of the figures for download, which is available to authorised users.

Mugdha Gokhale
mugdha.gokhale@merck.com

[1] Pharmacoepidemiology, Center for Observational & Real-World Evidence, Merck, 770 Sumneytown Pike, West Point, PA 19486, USA

[2] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA

[3] Department of Medicine, University of North Carolina, Chapel Hill, NC, USA

Introduction

Real-world evidence (RWE) remains one of the most enticing concepts in medicine, surrounded by much buzz. Recent developments, including the first-ever regulatory approval of label expansion of IBRANCE (palbociclib) for male breast cancer based on RWE, have brought in a new era in the applicability of RWE in healthcare [1]. RWE is defined as 'clinical evidence about the usage and potential benefits or risks of a medical product derived from analysing real-world data (RWD)' [2]. RWD are data relating to patient health status and/or the delivery of healthcare routinely collected from different sources and find application in many areas including therapeutic development to comparative effectiveness/safety, reimbursement, regulatory decision-making and clinical guideline development [2, 3].

The reflection of 'diverse real-world practices' enhances the appeal of RWD making it more relatable than data from RCTs. However, this very element that makes RWD attractive also makes it challenging to work with. Additionally, inaccurate application of methods and shortage of adequate methodological know-how potentially threaten the validity of RWD studies [4]. In this paper we discuss commonly encountered issues and recommend key methodological considerations and potential solutions in the design, implementation and evaluation of real-world pharmacoepidemiological studies. This paper provides a general overview of a broad topic and because a detailed discussion on each subtopic is beyond the scope of this review, we have cited several references in relevant sections for interested readers to explore further.

Defining the research question using a causal inference framework

It is a misconception that the entire purpose of RWD is to reduce the cost or complexity of trials (although this is feasible and done in some settings) or simply to get evidence 'without randomisation'. RWE when done correctly is an important stand-alone source of evidence that complements RCTs and laboratory and other studies, which together inform decision-making. Understanding the research question is crucial to ensure that the right tools to generate robust RWE are employed.

Researchers should accurately describe the goals of real-world studies using a causal inference framework, like they would for an RCT, including any nuances (e.g. are we comparing initiation of treatment A vs treatment B or are we comparing patients switching from treatment A to treatment B vs patients staying on treatment A?) [5]. While RWE is inherently relevant to clinical practice, it examines a different pattern of care (i.e. RWE and RCTs ask different questions). However, imagining an RWE study as a hypothetical trial forces a more stringent thought process about the intervention, comparators, timelines, outcomes and confounders [6, 7]. In estimating a causal effect, we ideally want to examine all potential outcomes in the same patients during the same time period, under contrasting treatments [8]. However, this is impossible, as for each patient the outcome can be observed only under one treatment. We therefore compare the outcomes of two groups: treatment A vs an 'exchangeable' substitute treatment B [9, 10]. The validity of effect estimates depends on how valid the substitution is [10]. While there are no guarantees that an effect estimate can be causally interpreted, setting a causal goal for the study lays the foundation for robust design and analytical decisions [5–8].

Data sources

Table 1 describes several data sources and provides examples of their application in diabetes research. RWD sources include administrative claims data [11–13], electronic health records [14, 15] and disease or treatment registries [16, 17]. Additionally, patient-generated data from surveys, questionnaires, smartphone apps and social media are increasingly being considered for the purposes of pharmacovigilance, patient characterisation and disease understanding [11, 18, 19]. However, these need careful evaluation as not all health apps are thoroughly validated and their pace of growth is fast outpacing the vetting process [20]. Data linkages with appropriate safeguards offer opportunities to combine useful features from two or more data sources to conduct real-world studies [21].

Study design

Several classification schemes exist for real-world study designs [22] but, broadly, cohort studies, case–control studies and self-controlled case series are the three basic types [23]. Cohort studies follow patients from the treatment to the outcome. Case–control studies select disease cases and controls from the source population and compare treatment histories in the two groups, thereby introducing several avenues for biases. Cohort design is a direct analogue of an experiment and has generally become the standard design for observational studies unless an alternate design is dictated by the research question. Self-controlled methods compare treatments and outcomes within the same individual rather than across groups of individuals by looking at different treatment times within the same person. They are a good fit in settings with acute recurrent or non-recurrent events and intermittent exposures and transient effects, assuming availability of precise timings [24]. Choice of a study design depends on several factors including the research question of interest, rarity of the exposure/outcome and avenues for biases. We direct interested readers to publications by Rothman et al [23] and Hallas and Pottegård [25] for further reading on this subject. As cohort studies are most intuitive when assessing incidence, natural history or comparative effectiveness/safety, we will focus much of our further discussion with this design in mind. Recently, clear design diagrams have been proposed to intuitively visualise studies conducted using healthcare databases/electronic medical records (Fig. 1) [26].

Potential biases

The biggest criticism of real-world studies is their potential for systematic error (biases). These are broadly classified as confounding (due to lack of randomisation), selection bias (due to procedures used to select study population) and information bias (measurement error) [23].

Confounding Confounding is the distortion of the treatment–outcome association when the groups being compared differ with respect to variables that influence the outcome. In

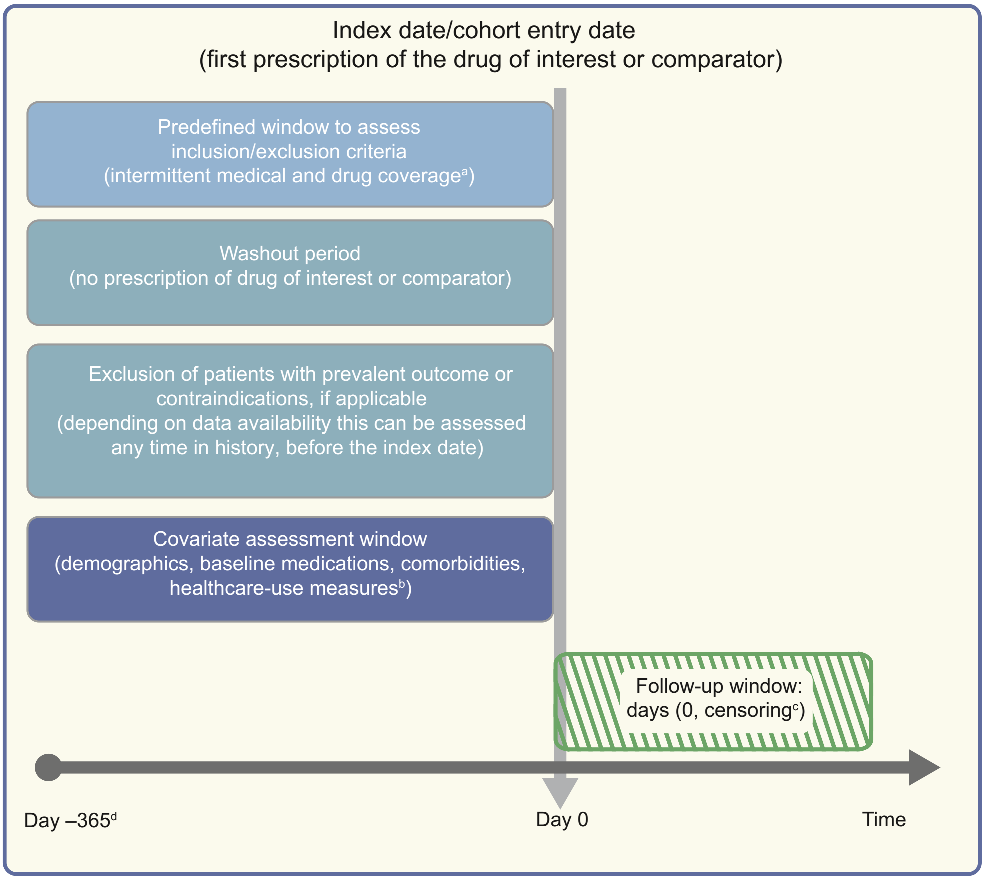
Table 1 Examples of RWD sources and applications to diabetes research

RWD source	Merits	Caveats	Potential areas for application in diabetes research
Administrative claims data: Insurance claims for pharmacy prescriptions and medical inpatient and outpatient visits submitted for billing purposes by government or commercial payers Include cost information, date/place of service and patient demographics, all linked by a common patient identifier	Longitudinally follow patients as they navigate through the healthcare system Reliable for studying important medical encounters, diagnoses and treatment using variables that are captured for reimbursement purposes Provide information on large samples of patients and their families, considered to be representative of the target population (commercially insured/populations under public health insurance programmes) Demographically and geographically diverse, relatively low cost and time-efficient vs RCTs	Primary purpose for data collection is administrative rather than for research Key clinical variables (e.g. severity), medications for which patients pay out-of-pocket, patient-reported outcomes, lifestyle variables and laboratory results are typically not captured Loss of follow-up, particularly in commercial claims data when patients switch employers/health plans (known censoring date due to availability of enrolment file) Identification of disease/treatment depends on accuracy of billing codes used and data require validation prior to use particularly for hard-to-diagnose rare conditions	Can be used in real-world studies to compare the effectiveness and safety of glucose-lowering therapies using active comparator new-user design [42, 59], patient characterisation, treatment utilisation [91] and health policy/cost [92] research, as well as burden of illness [93] studies Can be used to estimate basic prevalence or incidence measures of conditions within diabetes populations given large sample sizes and representativeness
EHR data: Data from patients' electronic medical records Data typically include information on medical diagnoses, procedures, medications, free text with physician notes, vital signs at each visit, laboratory results, clinical variables	Data collected to capture clinical care and contain rich data on clinical variables or other important confounders May provide rationale for treatment decisions depending on the quality of free text	Variability in the quality of data as clinical variables are often missing and may be recorded differently by different physicians Follow-up only available as long as patients remain in the healthcare system and seek care (unknown censoring date since no enrolment file) Typically, data from only one place of service are available and capture of information from other types of practices are often unreliable (e.g. in a general practice system, specialist data may not be accurately captured for all patients; hospitalisations for acute problems outside the system may not be captured)	Assessing comparative effectiveness or safety, treatment patterns and patient characterisation Typically less useful for cost assessments or prevalence/incidence estimation Analyses of EHR data have been shown to improve glycaemic control, reduce emergency department visits and non-elective hospitalisations [94, 95]
Patient-generated data: Data from surveys, questionnaires, smartphone apps and social media that allow continuous data capture Information is provided mainly by patients, rather than by providers	Questionnaire/survey data sources provide data on quality-of-life measures, which are hard to find in other data sources Can be used as external validation datasets May find particular relevance in pharmacovigilance, particularly rare adverse events associated with treatments, and factors predicting patients' adherence, behaviours and attitudes Some data include real-time monitoring to allow tracking of selected measures and symptoms	Use of these sources implies reliance on self-reported variables, leading to recall bias, selective reporting and missing data on important patient characteristics and medical variables Limited generalisability and internal validity, as the clinical outcomes reported are often not validated and authenticity is often unverifiable Utility only in specific settings after careful evaluation and vetting	The FDA-approved WellDoc BlueStar System is a healthcare app that provides secure capture of blood glucose data and aids in diabetes self-management [96]
Patient registries: Repositories of rich information on specific disease or treatment	Include data on patients' characteristics and medical variables, including rich clinical information on disease or treatments of interest Allow long patient follow-up	Validity greatly depends on what type of patients are selected into the registry (voluntary vs mandatory enrolment) Expensive to maintain May not contain information on other comorbidities or concurrent	The diabetes collaborative registry, organised by the leading societies in diabetes research, provides RWD on diabetes patient care and treatment [17]

Table 1 (continued)

RWD source	Merits	Caveats	Potential areas for application in diabetes research
	Useful in areas where richness of information related to a specific disease/treatment is desirable (e.g. rare tumours) and in unique populations (e.g. pregnancy registries)	treatment; more potential for missing data	
Data linkages: Data from two or more sources are linked to bring together the information needed, assuming appropriate safeguards are applied	Bring together data from disparate sources allowing capture of comprehensive information needed in a particular research setting (e.g. linking administrative claims with EHRs would enable combination of longitudinal follow-up, cost information that may be lacking in EHRs, with clinical variables that are incomplete in claims) Help minimise missing data on key variables, reducing misclassification	Validity of results depends on the quality of linkage Expensive to link and maintain linked data sources Challenges in linking data due to different purposes of data collection, discrepancies in data recording, legal/confidentiality issues	Several studies using linked data are being conducted in diabetes patients, predicting hospital admissions [97], cancer outcomes [98] and weight gain with diabetes treatments [99]

EHR, electronic health record



Index date/cohort entry date
(first prescription of the drug of interest or comparator)
Predefined window to assess inclusion/exclusion criteria (intermittent medical and drug coverage[a])
Washout period (no prescription of drug of interest or comparator)
Exclusion of patients with prevalent outcome or contraindications, if applicable (depending on data availability this can be assessed any time in history, before the index date)
Covariate assessment window (demographics, baseline medications, comorbidities, healthcare-use measures[b])
Follow-up window: days (0, censoring[c])
Day –365[d]
Day 0
Time

Fig. 1 Framework for a cohort study using an administrative claims or electronic medical record database, with methodology from Schneeweiss et al [26], and using templates from www.repeatinitiative.org/projects.html, which are licensed under a Creative Commons Attribution 3.0 Unported License. [a]Typically, a gap of up to 45 days in medical or pharmacy enrolment is allowed; [b]covariates are measured in the 6 month period before entry into the cohort, and demographics are measured on day zero; [c]earliest of outcome of interest, switching or discontinuation of study drugs, death, disenrolment, or end of the study period; [d]365 days pre-index are shown for illustrative purposes; this could be any predefined time before the index date deemed appropriate, and tailored to the study question at hand. This figure is available as part of a downloadable slideset

comparing drug treatments, confounding by indication is a common issue that occurs because patients have an 'indication' for a particular drug [27]. As an example, comparing patients prescribed insulin vs oral glucose-lowering agents leads to confounding by indication as the two populations are imbalanced on the 'indication' (severe vs milder diabetes). When a treatment is known to be associated with an adverse event, confounding by contraindication is possible. In comparing thiazolidinediones with dipeptidyl peptidase-4 (DPP-4) inhibitors to assess the risk for heart failure, for example, patients with existing heart conditions are likely to be channelled away from thiazolidinediones. Restricting the study population to those without prevalent heart failure will therefore minimise intractable confounding by contraindication [28].

In real-world studies confounding by frailty is possible. This is a particular problem in older adults, as frail, close-to-death patients are less likely to be treated with preventive treatments. Thus, when comparing users vs non-users of a particular drug to assess outcomes associated with frailty (e.g. mortality risk), the non-user group is likely to have higher mortality risk and make the drug look better than it really is [29, 30].

Selection bias This bias occurs when the selected population is not representative of the target population to which inference is to be drawn (due to selective survival rate, differential losses to follow-up, non-response, etc.) [31]. Selection bias is sometimes intertwined with confounding depending on the setting in which it occurs (e.g. epidemiologists sometimes use the term selection bias to mean 'confounding by indication', others use the term selection bias when confounders are unmeasured) [32].

Information bias This arises due to inaccurate measurement or misclassification of treatments, outcome or confounders [23]. Its effect on results depends on whether misclassification is differential or non-differential across the treatments being compared. In a cohort study, non-differential exposure misclassification occurs when treatment status is equally misclassified among patients who develop or do not develop the outcome. As an illustration, if 10% of patients in both treatment A and treatment B groups received free drug samples (and therefore no prescription record in claims data), an equal proportion of patients in each group will be misclassified as 'unexposed'. Non-differential outcome misclassification in a cohort study occurs when patients who develop the outcome are equally misclassified in treatment A and treatment B groups (e.g. 15% of healthy patients receiving treatment A or treatment B are misclassified as having lung cancer). Differential misclassification occurs when misclassification of treatment status is uneven between individuals that have or do not have the outcome, or when misclassification of the outcome is not the same between treatment A and treatment B. While non-differential misclassification of treatments and outcomes will generally bias estimates towards the null, differential misclassification can lead to spurious associations or can mask true effects. The effect of misclassification on results also depends on whether the results are reported as absolute or relative [23]. Absolute measures present difference in risk of the outcome in treatment A vs treatment B, while relative measures present the ratio of risk of outcome for treatment A vs treatment B. When misclassification is non-differential, studies reporting absolute measures should have outcome definitions with high sensitivity and specificity as low values for either can lead to bias. In studies reporting relative measures, near-perfect specificity, even at the cost of low sensitivity, is desired [33].

Another common criticism of RWD concerns 'missing data'. The commonly used strategy of excluding records with missing data can severely bias results. Multiple imputation methods for mitigating the effect of missing data have been shown to decrease bias and improve precision in a variety of disease areas including diabetes [34, 35]. Methods for addressing missing data should be based on a careful consideration of reasons for missingness and availability of validation datasets needed for imputation methods [36].

Time-related biases These are biases that misclassify person-time attributed to the treatment. Immortal time bias is one such bias arising from misclassification of the time before the treatment during which a patient, by design, could not have experienced the outcome and the patients have to be event-free until treatment starts [37]. Misclassifying this time or excluding it from the analysis altogether leads to immortal time bias [37, 38]. This is exacerbated in studies comparing treatment users vs non-users (Fig. 2) but can occur when comparing active treatments without careful consideration of person-time. Consider an example comparing a sulfonylurea vs metformin, where metformin users consisted of patients with or without prior sulfonylurea use. For the metformin patients with prior sulfonylurea use, their time on sulfonylureas before metformin was misclassified as 'metformin-exposed' which led to immortal time bias and spuriously suggested the protective effect of metformin on mortality risk, since they had to survive to switch to metformin [39, 40].

In studies assessing cancer outcomes, events occurring shortly after initiation may not be meaningfully attributed to the exposed period, particularly since carcinogenic exposures typically have long induction periods [41]. Not counting person-time and events during a predefined time-lag after drug initiation accounts for both these periods (Fig. 3). Similarly, it is unlikely that patients stop being 'at risk' on the day after drug discontinuation and a period of latency should also be considered to provide an opportunity to capture the outcome that was potentially present subclinically before treatment discontinuation [41]. As an example, a recent study exploring the incidence of breast cancer with insulin glargine vs intermediate-acting insulin used induction and lag periods to account for cancer latency [42].

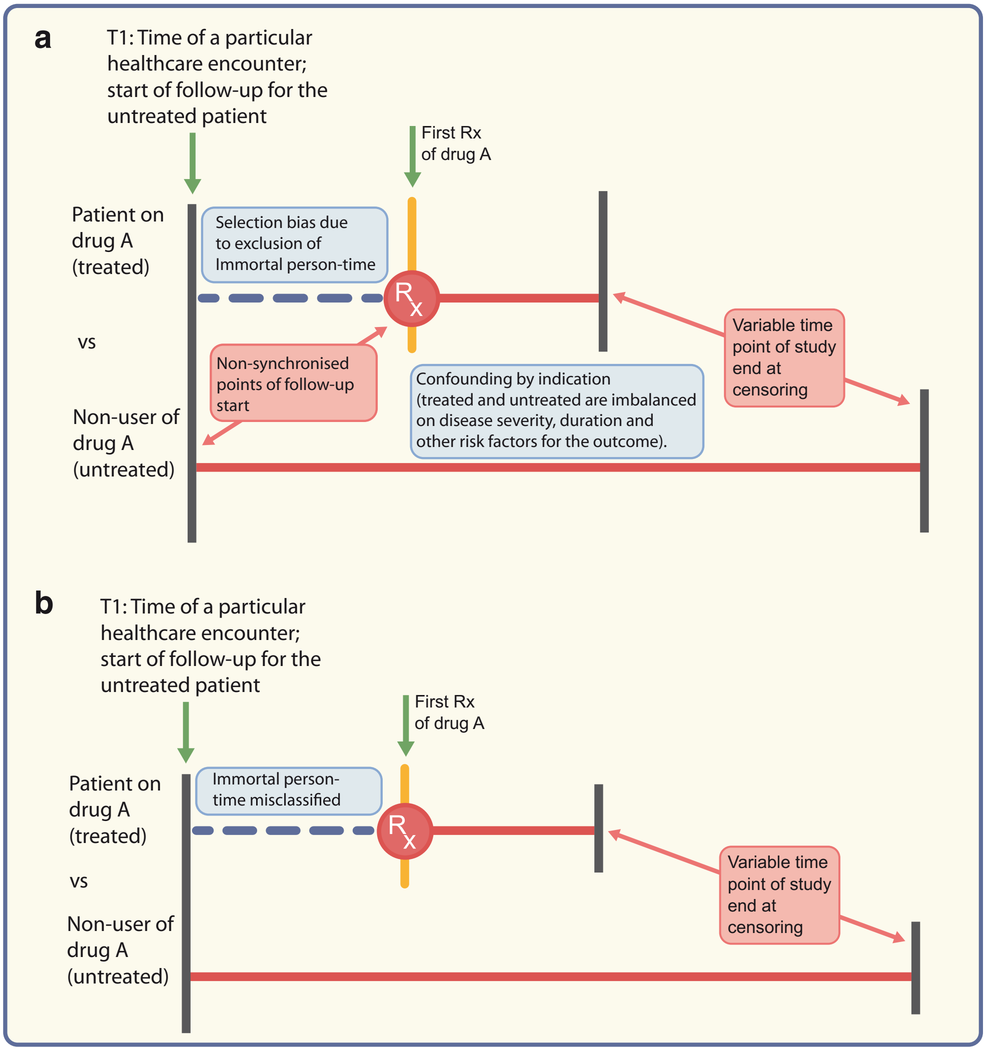

a
T1: Time of a particular healthcare encounter; start of follow-up for the untreated patient
First Rx of drug A
Patient on drug A (treated)
Selection bias due to exclusion of Immortal person-time
Rx
vs
Non-synchronised points of follow-up start
Confounding by indication (treated and untreated are imbalanced on disease severity, duration and other risk factors for the outcome).
Variable time point of study end at censoring
Non-user of drug A (untreated)
b
T1: Time of a particular healthcare encounter; start of follow-up for the untreated patient
First Rx of drug A
Patient on drug A (treated)
Immortal person-time misclassified
Rx
vs
Variable time point of study end at censoring
Non-user of drug A (untreated)

Fig. 2 Depiction of problems encountered when comparing treated vs untreated (not using active comparator) patients; drug A could, as an example, be a DPP-4 inhibitor. (**a**) Different times of follow-up (starting at the initiation date for the treated patients or time of healthcare encounter T1 for the untreated patients) will lead to selection bias if immortal person-time is excluded from the analysis. Confounding by indication may arise from the imbalance between the two groups on 'indication'. (**b**) Even if the follow-up for both groups starts from time T1, the time between T1 and drug initiation would be misclassified as 'time on drug A' when in reality the patient was not on drug A before the first prescription. Red horizontal lines represent study timeline. Rx, prescription. This figure is available as part of a downloadable slideset

Prevalent-user biases Prevalent users are patients already on treatment before follow-up starts and therefore more tolerant of the drug. Methodological problems due to inclusion of prevalent users are illustrated by inconsistent results from studies examining cancer incidence with insulin glargine, depending on the study design used [43, 44]. However, the most striking and frequently cited example illustrating these issues is a series of studies highlighting the discrepancies in the estimated effects of hormone therapy on cardiovascular outcomes between new users and prevalent users in the same data [45–48]. The Nurses' Health cohort study reported a decreased risk of major CHD in prevalent users of oestrogen and progestin compared with non-users, in contrast to results from the RCT which showed an increased risk in the oestrogen + progestin arm relative to placebo [49, 50]. A re-analysis of the Nurses' Health study cohort comparing new users of hormone therapy vs non-initiators demonstrated results in line with the RCT, highlighting the issues due to inclusion of prevalent users [51]. As the prevalent users have 'survived' treatment, any patients who experienced early events (susceptible patients) will be automatically excluded in prevalent-user studies, introducing substantial bias if the hazard for the outcome varies depending on time spent on treatment [48, 52]. In studies with a mix of prevalent and incident users, the differential proportion of prevalent users across two groups being compared leads to selection bias and also obscures early events if the prevalent users contribute more person-time. Moreover, since confounders are affected by prior treatment, they are mediators in a causal pathway between the treatment and outcome, and any analytical adjustment would worsen the bias [48].

Methods for minimising bias by study design and analysis

Active comparator new-user design

To avoid prevalent-user biases, new-user design has been recommended as almost a default strategy, except in settings

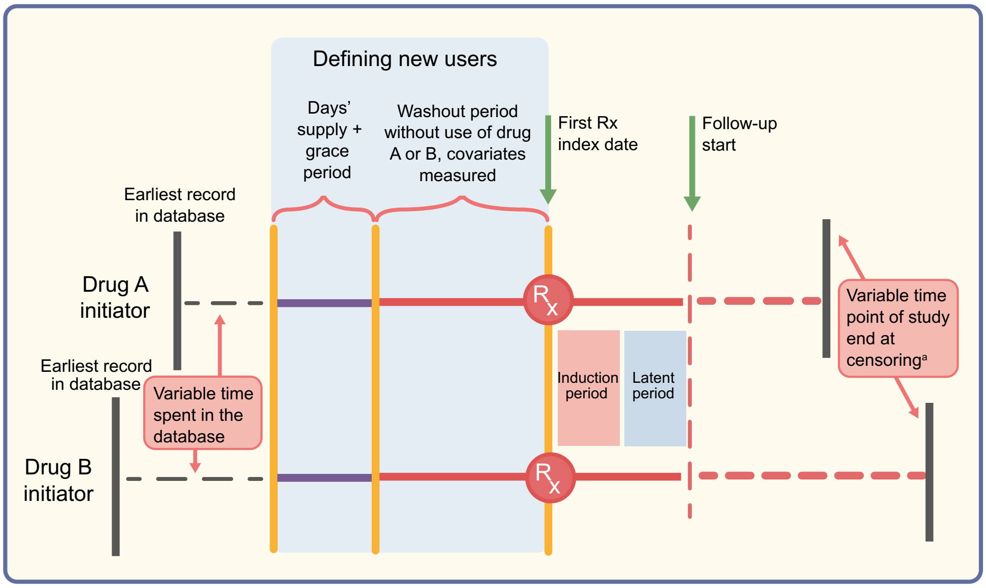

Defining new users
Days' supply + grace period
Washout period without use of drug A or B, covariates measured
First Rx index date
Follow-up start
Earliest record in database
Drug A initiator
Rx
Variable time point of study end at censoring[a]
Earliest record in database
Variable time spent in the database
Induction period
Latent period
Drug B initiator
Rx

Fig. 3 Schematic diagram of the active comparator new-user design comparing two glucose-lowering drugs. The top and bottom horizontal lines represent study timelines for initiators of drug A (e.g. DPP-4 inhibitor) and a therapeutically equivalent drug B (e.g. pioglitazone), respectively. Both groups of patients spend a variable amount of time in the database before 'new use' is defined. To define the new-use period for both groups, we need a predefined period equal to 'expected days' supply plus the grace period' without a prescription being filled for treatment A or treatment B (indicated by solid purple lines) prior to the start of the washout period (solid orange lines). The washout period should also be free of any prescriptions for A or B and covariates are measured during this time. The index date indicates the date of the first prescription (Rx) and is followed by induction and latent periods during which person-time and outcomes are not counted. Solid red lines represent this time after the first prescription. Follow-up, indicated by the dashed red lines, starts at the end of the latent period and ends at censoring. Note that patients' timelines for start of follow-up are intuitively synchronised by the active comparator new-user design, even though the patients can have variable start points of enrolment in the database or variable time points for end of follow-up/study end. [a]If censoring is due to drug discontinuation, a predefined lag period should be considered before stopping to count outcomes and person-time. This figure is available as part of a downloadable slideset

where inclusion of prevalent users is preferable (e.g. describing burden of illness) [53]. This design includes initiators of a drug, after a washout period without that drug (treatment-naivety not necessary) and provides an intuitive timeline to start follow-up [48]. Because new-user designs restrict the population to drug initiators, concerns have been expressed about reduced generalisability and precision at the cost of high internal validity. Modifications of the new-user design, such as the prevalent new-user designs, have recently been proposed to address this (e.g. while comparing new-to-market vs older drugs) [54]. Such designs propose including patients switching from older to newer drugs to increase precision. However, precision gain needs to be carefully weighed against the mixing of research questions (initiation vs switching) and the potential for biases introduced by comparing switchers with patients who remain on treatment [55].

The merits of a new-user design are further amplified by comparing drugs in clinical equipoise [56]. Comparing treated and untreated patients opens the door to a host of biases (Fig. 2), which can be overcome by the active comparator new-user design comparing new users of therapeutically equivalent drugs (active comparators; Fig. 3). This design makes the two cohorts 'exchangeable' with respect to baseline disease severity and outcome risk and the follow-up can start from an intuitive, synchronised time point [57, 58]. The demonstrated balance of measured characteristics may also increase the probability of balance of unmeasured covariates, although this cannot be empirically demonstrated [28, 59]. Often there may be situations in diabetes research where an active comparator is not available (e.g. an RWE study emulating a placebo-controlled trial). In such cases, synchronising cohorts based on any healthcare encounters that make the two cohorts as substitutable as possible is still preferred over comparing with non-users [57]. In a recent example illustrating this principle, the risk of cardiovascular outcomes with the sodium–glucose cotransporter-2 (SGLT2) inhibitor canagliflozin was assessed relative to non-SGLT2-inhibitors rather than to non-users of canagliflozin (which could have led to inclusion of diabetes patients not on pharmacological therapy and therefore caused imbalance of patient characteristics) [60].

The active comparator new-user design is analogous to a head-to-head RCT comparing two drugs in equipoise. It allows following patients by ignoring treatment changes over time, analogous to the 'intent-to-treat' analyses in RCTs. This may introduce treatment misclassification bias towards the null and should be avoided, particularly in studies assessing harm, to avoid masking actual treatment-associated harm. Another option is the 'as-treated' approach where follow-up is censored at treatment discontinuation, switching or augmentation. A caveat with the 'as-treated' approach is potential selection bias introduced because of informative censoring (i.e. patients censored because they made treatment changes are not representative of patients who remain on treatment). This needs to be addressed in the analysis using inverse probability of censoring weights [32].

Recent applications of these designs in diabetes research include new-user studies on SGLT2 inhibitors demonstrating no increased risk of amputations relative to non-SGLT2 inhibitors, but increased risk relative to the most appropriate DPP-4 inhibitor comparator using restrictive study criteria and robust analytic techniques [60, 61].

Analysis

An example that naturally fits with the active comparator new-user study is the use of propensity scores, a powerful tool for controlling measured confounding [62–66]. A propensity score is a summary score estimating the probability of treatment A vs treatment B based on patients' baseline characteristics. Once estimated, propensity scores can be implemented by matching, weighting and stratification on the score [62, 63, 67, 68], all of which allow empirical demonstration of covariate balance before and after implementation. Propensity scores can also be included in an outcome model, although this takes away the ability to empirically 'see' the adjustment and has other disadvantages so is therefore discouraged [67]. The choice of method used to implement propensity scores (matching, stratification, different types of weighting) depends on the target population to which inference needs to be drawn and the extent of unmeasured residual confounding [62, 69–74].

Other methods such as disease risk scores (summary scores based on baseline outcome risk) can have advantages in specific settings [75, 76]. When substantial unmeasured confounding is expected, instrumental variable methods might be used to obtain unbiased effect estimates [77, 78]. Instrumental variables are variables that affect treatment choice but not the outcome of interest other than through treatment. Examples include physicians' preference and any rapid change in treatments (e.g. due to market access, guideline changes, warnings about serious side effects). All of these methods, however, are based on a number of assumptions that should be evaluated when conducting and interpreting real-world studies. Further, more than one method can be considered as supplementary sensitivity analyses after clearly specifying the reasons a priori. Given the dynamic treatment patterns in routine clinical practice (discontinuation, re-initiation, switching treatment, etc.), analyses often need to account for time-varying treatments and confounders depending on the research question of interest.

Recently, the utility of machine learning in causal inference has been explored [79, 80]. Machine learning algorithms have been shown to perform well in estimating propensity scores in certain settings by reducing model mis-specification [81, 82] but can amplify bias in certain settings (e.g. if use of instrumental variables in propensity score estimation is encouraged) [83]. A concern with these methods is the lack of easy interpretability and the risk of being data-driven rather than being informed by substantive knowledge and therefore need careful consideration before being used.

Newer avenues for applicability of RWE

RWD are increasingly being used to predict outcomes from clinical trials, which supports efficient resource management, faster drug approval times and making medicines available sooner for patients. A recent example is a study comparing linagliptin vs glimepiride using RWD from Medicare and commercial data [84]. While this study demonstrated linagliptin's 'non-inferiority' in line with the findings of the CAROLINA trial, the magnitude was smaller than that observed in the trial. This was likely due to differences in the nature of treatments being compared rather than a lack of robust methodology. Despite the difference in magnitude of results, this supports the value that RWD brings. Another application is the use of an RWD-based comparator for single-arm trials when using a randomised control arm is not feasible, as was done with BLINCYTO (blinatumomab) indicated for leukaemia treatment [2]. We use blinatumomab as a powerful example of use of RWD in regulatory decision-making. It is not inconceivable that RWD may find applications in the future in areas where randomised trials may be deemed unethical, such as treatment of heart failure without background SGLT2 inhibitor therapy, or for the prevention of rare events where sample sizes could become prohibitive. The main challenge to address here is the differences between trial participants vs patients in routine practice, including the potential for differential recording of characteristics, warranting deeper design and analytical considerations depending on the nature and extent of differences. Efforts are also ongoing to map the potential effects of RCT data to real-world populations, although we are not aware of examples of this in diabetes research. Pragmatic trials (that measure effectiveness of the treatment/intervention in routine clinical practice) including RWD are also increasingly being explored in a number of disease areas [85, 86]. Finally, we may be at the verge of a paradigm shift with respect to classification of validity and hierarchy of study designs. The approach of prioritising internal validity (getting unbiased estimates) at the cost of external validity (generalisability or transportability of results), and our current thinking of internal validity as a prerequisite for external validity can negatively affect the value that RWE brings. Westreich et al recently proposed a joint measure of the validity of effect estimates (target validity) and defined target bias as any deviation of the estimated treatment effect from the true treatment effect in a target population rather than the current distinction between internal and external validity [87].

Conclusion

The value of RWE lies in going beyond the constraints of RCTs to understand the effects in real-world populations. However, the hopes of 'quick wins' with RWE need to be balanced with a knowledge of robust methodology. We have focused our discussion on key concepts, methods and recommendations in the hope that readers are better informed of the utility and limitations of a particular RWE study that they encounter. The following key points should be looked for when evaluating an RWE study: a clearly articulated research question; a fit-for-purpose data source; a state-of-the-art design including appropriate comparators; covariate balance; analysis methods including sensitivity analyses; and the likelihood of being able to reasonably replicate the study in another similar setting. Several guidelines have come into existence to assist investigators with proper conduct, interpretation and reporting of real-world studies [88–90]. As the field continues to grow, it is important for scientific journals and regulatory agencies to use peer reviewers with adequate methodological know-how to ensure dissemination of high-quality RWE and maximise its utility in decision-making.

Authors' relationships and activities MG was a full-time employee of GlaxoSmithKline at the time of drafting this manuscript. She is now a full-time employee of and owns stocks in Merck and has no potential conflicts of interest relevant to this article. JBB's contracted consulting fees and travel support for contracted activities are paid to the University of North Carolina by Adocia, AstraZeneca, Dance Biopharm, Dexcom, Eli Lilly, Fractyl, GI Dynamics, Intarcia Therapeutics, Lexicon, MannKind, Metavention, NovaTarg, Novo Nordisk, Orexigen, PhaseBio, Sanofi, Senseonics, vTv Therapeutics and Zafgen. JBB reports grant support from AstraZeneca, Eli Lilly, Intarcia Therapeutics, Johnson & Johnson, Lexicon, Medtronic, Novo Nordisk, Sanofi, Theracos, Tolerion and vTv Therapeutics and is a consultant to Cirius Therapeutics Inc., CSL Behring, Mellitus Health, Neurimmune AG, Pendulum Therapeutics and Stability Health. JBB holds stock/options in Mellitus Health, Pendulum Therapeutics, PhaseBio and Stability Health and is supported by grants from the National Institutes of Health (UL1TR002489, U01DK098246, UC4DK108612, U54DK118612, P30DK124723), PCORI and ADA. TS receives investigator-initiated research funding and support from the National Institute on Aging as a principal investigator (R01 AG056479) and from the National Institutes of Health as co-investigator (R01 CA174453, R01 HL118255 and R21-HD080214). He also receives salary support as co-Director of the Biostatistics, Epidemiology, and Research Design (BERD), North Carolina Translational and Clinical Sciences Institute (UL1TR002489) and from the Center for Pharmacoepidemiology (current members are GlaxoSmithKline, UCB BioSciences, Merck and Takeda). TS also receives research support from pharmaceutical companies (Amgen, AstraZeneca and Novo Nordisk), and from a generous contribution from N. A. Dreyer to the Department of Epidemiology, University of North Carolina at Chapel Hill. TS does not accept personal compensation of any kind from any pharmaceutical company. He owns stock in Novartis, Roche, BASF, AstraZeneca and Novo Nordisk. No other potential conflicts of interest relevant to this article were reported.

Contribution statement All authors made substantial contributions to the conception of this manuscript, drafting the article and revising it critically for important intellectual content. All authors approved the version to be published.

References

1. Wedam S, Fashoyin-Aje L, Bloomquist E et al (2020) FDA approval summary: palbociclib for male patients with metastatic breast cancer. Clin Cancer Res 26(6):1208–1212. https://doi.org/10.1158/1078-0432.CCR-19-2580
2. U.S. FDA (2019) Webinar: framework for FDA's Real-World Evidence Program – Mar 15, 2019. Available from https://www.fda.gov/drugs/webinar-framework-fdas-real-world-evidence-program-mar-15-2019. Accessed 20 Dec 2019
3. U.S. FDA (2019) Submitting documents using real-world data and real-world evidence to FDA for drugs and biologics guidance for industry - May 9, 2019. Available from https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drugs-and-biologics-guidance. Accessed 20 Dec 2019
4. Allie Nawrat, Pharma Technology Focus (2019) Real world data - how can it improve clinical trial outcomes. Available from https://www.pharmaceutical-technology.com/features/real-world-data-improving-outcomes/. Accessed 20 Dec 2019
5. Hernan MA (2018) The C-word: scientific euphemisms do not improve causal inference from observational data. Am J Public Health 108(5):616–619. https://doi.org/10.2105/AJPH.2018.304337
6. Hernan MA, Robins JM (2016) Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 183(8):758–764. https://doi.org/10.1093/aje/kwv254
7. Hernan MA (2005) Invited commentary: hypothetical interventions to define causal effects–afterthought or prerequisite? Am J Epidemiol 162(7):618–620; discussion 621-612. https://doi.org/10.1093/aje/kwi255
8. Rubin DB (2005) Causal inference using potential outcomes: design, modeling, decisions. J Am Stat Assoc 100(469):322–331. https://doi.org/10.1198/016214504000001880
9. Nørgaard M, Ehrenstein V, Vandenbroucke JP (2017) Confounding in observational studies based on large health care databases: problems and potential solutions–a primer for the clinician. Clin Epidemiol 9:185–193. https://doi.org/10.2147/CLEP.S129879
10. Maldonado G, Greenland S (2002) Estimating causal effects. Int J Epidemiol 31(2):422–429. https://doi.org/10.1093/ije/31.2.422
11. Nabhan C, Klink A, Prasad V (2019) Real-world evidence-what does it really mean? JAMA Oncol 5(6):781–783. https://doi.org/10.1001/jamaoncol.2019.0450
12. Sturmer T, Jonsson Funk M, Poole C, Brookhart MA (2011) Nonexperimental comparative effectiveness research using linked healthcare databases. Epidimiology 22(3):298–301. https://doi.org/10.1097/EDE.0b013e318212640c
13. Strom BL (2001) Data validity issues in using claims data. Pharmacoepidemiol Drug Saf 10(5):389–392. https://doi.org/10.1002/pds.610

14. Casey JA, Schwartz BS, Stewart WF, Adler NE (2016) Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 37(1):61–81. https://doi.org/10.1146/annurev-publhealth-032315-021353
15. Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L (2018) Promises and pitfalls of electronic health record analysis. Diabetologia 61(6):1241–1248. https://doi.org/10.1007/s00125-017-4518-6
16. Nelson EC, Dixon-Woods M, Batalden PB et al (2016) Patient focused registries can improve health, care, and science. BMJ 354:i3319. https://doi.org/10.1136/bmj.i3319
17. Diabetes Collaborative Registries (2019) The Diabetes Collaborative Registry. Transforming the future of diabetes care. Available from https://cvquality.acc.org/NCDR-Home/registries/outpatient-registries/the-diabetes-collaborative-registry. Accessed 20 Dec 2019
18. McDonald L, Malcolm B, Ramagopalan S, Syrad H (2019) Real-world data and the patient perspective: the PROmise of social media? BMC Med 17(1):11. https://doi.org/10.1186/s12916-018-1247-8
19. Pierce CE, Bouri K, Pamer C et al (2017) Evaluation of facebook and twitter monitoring to detect safety signals for medical products: an analysis of recent FDA safety alerts. Drug Saf 40(4):317–331. https://doi.org/10.1007/s40264-016-0491-0
20. Kuehn BM (2015) Is there an app to solve app overload? JAMA 313(14):1405–1407. https://doi.org/10.1001/jama.2015.2381
21. Rivera DR, Gokhale MN, Reynolds MW et al (2020) Linking electronic health data in pharmacoepidemiology: appropriateness and feasibility. Pharmacoepidemiol Drug Saf 29(1):18–29. https://doi.org/10.1002/pds.4918
22. Pearce N (2012) Classification of epidemiological study designs. Int J Epidemiol 41(2):393–397. https://doi.org/10.1093/ije/dys049
23. Rothman KJ, Greenland GS, Lash TL (2008) Modern epidemiology, 3rd edn. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadephlia
24. Petersen I, Douglas I, Whitaker H (2016) Self controlled case series methods: an alternative to standard epidemiological study designs. BMJ 354:i4515
25. Hallas J, Pottegård A (2014) Use of self-controlled designs in pharmacoepidemiology. J Intern Med 275(6):581–589. https://doi.org/10.1111/joim.12186
26. Schneeweiss S, Rassen JA, Brown JS et al (2019) Graphical depiction of longitudinal study designs in health care databases. Ann Intern Med 170(6):398–406. https://doi.org/10.7326/M18-3079
27. Blais L, Ernst P, Suissa S (1996) Confounding by indication and channeling over time: the risks of β_2-agonists. Am J Epidemiol 144(12):1161–1169. https://doi.org/10.1093/oxfordjournals.aje.a008895
28. Gokhale M, Buse JB, Jonsson Funk M et al (2017) No increased risk of cardiovascular events in older adults initiating dipeptidyl peptidase-4 inhibitors vs therapeutic alternatives. Diabetes Obes Metab 19(7):970–978. https://doi.org/10.1111/dom.12906
29. Glynn RJ, Knight EL, Levin R, Avorn J (2001) Paradoxical relations of drug treatment with mortality in older persons. Epidemiology 12(6):682–689. https://doi.org/10.1097/00001648-200111000-00017
30. Zhang H, McGrath L, Ellis A, Wyss R, Lund J, Stürmer T (2019) Restriction of pharmacoepidemiologic cohorts to initiators of unrelated preventive drug classes to reduce confounding by frailty in older adults. Am J Epidemiol 188(7):1371–1382. https://doi.org/10.1093/aje/kwz083
31. Coggon D, Barker D, Rose G (2009) Epidemiology for the uninitiated, 5th edn. Wiley, New York
32. Hernán MA, Hernández-Díaz S, Robins JM (2004) A structural approach to selection bias. Epidimiology 15(5):615–625. https://doi.org/10.1097/01.ede.0000135174.63482.43
33. Chubak J, Pocobelli G, Weiss NS (2012) Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol 65(3):343–349. e342. https://doi.org/10.1016/j.jclinepi.2011.09.002
34. Read SH, Lewis SC, Halbesma N, Wild SH (2017) Measuring the association between body mass index and all-cause mortality in the presence of missing data: analyses from the Scottish National Diabetes Register. Am J Epidemiol 185(8):641–649. https://doi.org/10.1093/aje/kww162
35. Harel O, Mitchell EM, Perkins NJ et al (2018) Multiple imputation for incomplete data in epidemiologic studies. Am J Epidemiol 187(3):576–584. https://doi.org/10.1093/aje/kwx349
36. Hughes RA, Heron J, Sterne JAC, Tilling K (2019) Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol 48(4):1294–1304. https://doi.org/10.1093/ije/dyz032
37. Suissa S (2007) Immortal time bias in pharmacoepidemiology. Am J Epidemiol 167(4):492–499. https://doi.org/10.1093/aje/kwm324
38. Lévesque LE, Hanley JA, Kezouh A, Suissa S (2010) Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ 340:b5087. https://doi.org/10.1136/bmj.b5087
39. Bowker SL, Majumdar SR, Veugelers P, Johnson JA (2006) Increased cancer-related mortality for patients with type 2 diabetes who use sulfonylureas or insulin. Diabetes Care 29(2):254–258. https://doi.org/10.2337/diacare.29.02.06.dc05-1558
40. Suissa S, Azoulay L (2012) Metformin and the risk of cancer: time-related biases in observational studies. Diabetes Care 35(12):2665–2673. https://doi.org/10.2337/dc12-0788
41. Pottegard A, Friis S, Sturmer T, Hallas J, Bahmanyar S (2018) Considerations for pharmacoepidemiological studies of drug-cancer associations. Basic Clin Pharmacol Toxicol 122(5):451–459. https://doi.org/10.1111/bcpt.12946
42. Stürmer T, Marquis MA, Zhou H et al (2013) Cancer incidence among those initiating insulin therapy with glargine versus human NPH insulin. Diabetes Care 36(11):3517–3525. https://doi.org/10.2337/dc13-0263
43. Habel LA, Danforth KN, Quesenberry CP et al (2013) Cohort study of insulin glargine and risk of breast, prostate, and colorectal cancer among patients with diabetes. Diabetes Care 36(12):3953–3960. https://doi.org/10.2337/dc13-0140
44. Bradley MC, Chillarige Y, Lee H et al (2020) Similar breast cancer risk in women older than 65 years initiating glargine, detemir, and NPH Insulins. Diabetes Care 43(4):785–792. https://doi.org/10.2337/dc19-0614
45. Agency for Healthcare Research and Quality (2012) The incident user design in comparative effectiveness research. Available from https://effectivehealthcare.ahrq.gov/products/incident-user-design/research. Accessed 13 Dec 2019
46. Johnson ES, Bartman BA, Briesacher BA et al (2013) The incident user design in comparative effectiveness research. Pharmacoepidemiol Drug Saf 22(1):1–6. https://doi.org/10.1002/pds.3334
47. Petitti DB, Freedman DA (2005) Invited commentary: how far can epidemiologists get with statistical adjustment? Am J Epidemiol 162(5):415–418. https://doi.org/10.1093/aje/kwi224
48. Ray WA (2003) Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 158(9):915–920. https://doi.org/10.1093/aje/kwg231
49. Grodstein F, Stampfer MJ, Manson JE et al (1996) Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. N Engl J Med 335(7):453–461. https://doi.org/10.1056/NEJM199608153350701
50. Manson JE, Hsia J, Johnson KC et al (2003) Estrogen plus progestin and the risk of coronary heart disease. N Engl J Med 349(6):523–534. https://doi.org/10.1056/NEJMoa030808

51. Hernán MA, Alonso A, Logan R et al (2008) Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidimiology 19(6):766–779. https://doi.org/10.1097/EDE.0b013e3181875e61
52. Yola M, Lucien A (1994) Evidence of the depletion of susceptibles effect in non-experimental pharmacoepidemiologic research. J Clin Epidemiol 47(7):731–737. https://doi.org/10.1016/0895-4356(94)90170-8
53. Cox E, Martin BC, Van Staa T, Garbe E, Siebert U, Johnson ML (2009) Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II. Value Health 12(8):1053–1061. https://doi.org/10.1111/j.1524-4733.2009.00601.x
54. Suissa S, Moodie EE, Dell Aniello S (2017) Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores. Pharmacoepidemiol Drug Saf 26(4):459–468. https://doi.org/10.1002/pds.4107
55. Garry E, Buse JB, Gokhale M, Lund JL, Pate V, Sturmer T (2018) Implementation of the prevalent new user study design in the US Medicare population: benefit versus harm. Pharmacoepidemiol Drug Saf 27:167–167
56. Kramer MS, Lane DA, Hutchinson TA (1987) Analgesic use, blood dyscrasias, and case-control pharmacoepidemiology: a critique of the International Agranulocytosis and Aplastic Anemia Study. J Chronic Dis 40(12):1073–1081. https://doi.org/10.1016/0021-9681(87)90073-7
57. Lund JL, Richardson DB, Stürmer T (2015) The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. Curr Epidemiol Rep 2(4):221–228. https://doi.org/10.1007/s40471-015-0053-5
58. Yoshida K, Solomon DH, Kim SC (2015) Active-comparator design and new-user design in observational studies. Nat Rev Rheumatol 11(7):437–441. https://doi.org/10.1038/nrrheum.2015.30
59. Gokhale M, Buse JB, Gray CL, Pate V, Marquis MA, Stürmer T (2014) Dipeptidyl-peptidase-4 inhibitors and pancreatic cancer: a cohort study. Diabetes Obes Metab 16(12):1247–1256. https://doi.org/10.1111/dom.12379
60. Ryan PB, Buse JB, Schuemie MJ et al (2018) Comparative effectiveness of canagliflozin, SGLT2 inhibitors and non-SGLT2 inhibitors on the risk of hospitalization for heart failure and amputation in patients with type 2 diabetes mellitus: a real-world meta-analysis of 4 observational databases (OBSERVE-4D). Diabetes Obes Metab 20(11):2585–2597. https://doi.org/10.1111/dom.13424
61. Yang JY, Wang T, Pate V et al (2019) Sodium-glucose co-transporter-2 inhibitor use and risk of lower-extremity amputation: evolving questions, evolving answers. Diabetes Obes Metab 21(5):1223–1236. https://doi.org/10.1111/dom.13647
62. Stürmer T, Wyss R, Glynn RJ, Brookhart MA (2014) Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. J Intern Med 275(6):570–580. https://doi.org/10.1111/joim.12197
63. Brookhart MA, Wyss R, Layton JB, Stürmer T (2013) Propensity score methods for confounding control in nonexperimental research. Circ Cardiovasc Qual Outcomes 6(5):604–611. https://doi.org/10.1161/CIRCOUTCOMES.113.000359
64. Glynn RJ, Schneeweiss S, Stürmer T (2006) Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol 98(3):253–259. https://doi.org/10.1111/j.1742-7843.2006.pto_293.x
65. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S (2006) A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol 59(5):437–447. https://doi.org/10.1016/j.jclinepi.2005.07.004
66. Winkelmayer WC, Kurth T (2004) Propensity scores: help or hype? Nephrol Dial Transplant 19(7):1671–1673. https://doi.org/10.1093/ndt/gfh104
67. Stürmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ (2005) Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. Am J Epidemiol 161(9):891–898. https://doi.org/10.1093/aje/kwi106
68. Desai RJ, Rothman KJ, Bateman BT, Hernandez-Diaz S, Huybrechts KF (2017) A Propensity score based fine stratification approach for confounding adjustment when exposure is infrequent. Epidimiology 28(2):249–257. https://doi.org/10.1097/EDE.0000000000000595
69. Sato T, Matsuyama Y (2003) Marginal structural models as a tool for standardization. Epidimiology 14(6):680–686. https://doi.org/10.1097/01.EDE.0000081989.82616.7d
70. Yoshida K, Hernández-Díaz S, Solomon DH et al (2017) Matching weights to simultaneously compare three treatment groups: comparison to three-way matching. Epidimiology 28(3):387–395. https://doi.org/10.1097/EDE.0000000000000627
71. Li F, Thomas LE, Li F (2018) Addressing extreme propensity scores via the overlap weights. Am J Epidemiol 188(1):250–257
72. Crump RK, Hotz VJ, Imbens GW, Mitnik OA (2009) Dealing with limited overlap in estimation of average treatment effects. Biometrika 96(1):187–199. https://doi.org/10.1093/biomet/asn055
73. Yoshida K, Solomon DH, Haneuse S et al (2018) Multinomial extension of propensity score trimming methods: a simulation study. Am J Epidemiol 188(3):609–616
74. Glynn RJ, Lunt M, Rothman KJ, Poole C, Schneeweiss S, Stürmer T (2019) Comparison of alternative approaches to trim subjects in the tails of the propensity score distribution. Pharmacoepidemiol Drug Saf 28(10):1290–1298. https://doi.org/10.1002/pds.4846
75. Arbogast PG, Ray WA (2009) Use of disease risk scores in pharmacoepidemiologic studies. Stat Methods Med Res 18(1):67–80. https://doi.org/10.1177/0962280208092347
76. Glynn RJ, Gagne JJ, Schneeweiss S (2012) Role of disease risk scores in comparative effectiveness research with emerging therapies. Pharmacoepidemiol Drug Saf 21(Suppl 2):138–147. https://doi.org/10.1002/pds.3231
77. Brookhart MA, Rassen JA, Schneeweiss S (2010) Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf 19(6):537–554. https://doi.org/10.1002/pds.1908
78. Ertefaie A, Small DS, Flory JH, Hennessy S (2017) A tutorial on the use of instrumental variables in pharmacoepidemiology. Pharmacoepidemiol Drug Saf 26(4):357–367. https://doi.org/10.1002/pds.4158
79. Blakely T, Lynch J, Simons K, Bentley R, Rose S (2019) Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. Int J Epidemiol. https://doi.org/10.1093/ije/dyz132
80. Wyss R, Schneeweiss S, van der Laan M, Lendle SD, Ju C, Franklin JM (2018) Using super learner prediction modeling to improve high-dimensional propensity score estimation. Epidimiology 29(1):96–106. https://doi.org/10.1097/EDE.0000000000000762
81. Westreich D, Lessler J, Funk MJ (2010) Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J Clin Epidemiol 63(8):826–833. https://doi.org/10.1016/j.jclinepi.2009.11.020

82. Bi Q, Goodman KE, Kaminsky J, Lessler J (2019) What is machine learning? A primer for the epidemiologist. Am J Epidemiol 188(12):2222–2239. https://doi.org/10.1093/aje/kwz189
83. Myers JA, Rassen JA, Gagne JJ et al (2011) Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol 174(11):1213–1222. https://doi.org/10.1093/aje/kwr364
84. Patorno E, Schneeweiss S, Gopalakrishnan C, Martin D, Franklin JM (2019) Using real-world data to predict findings of an ongoing phase IV cardiovascular outcome trial–cardiovascular safety of linagliptin vs. glimepiride. Diabetes Care 42(12):2204–2210. https://doi.org/10.2337/dc19-0069
85. Dal-Ré R, Janiaud P, Ioannidis JPA (2018) Real-world evidence: how pragmatic are randomized controlled trials labeled as pragmatic? BMC Med 16(1):49. https://doi.org/10.1186/s12916-018-1038-2
86. Zuidgeest MGP, Goetz I, Groenwold RHH, Irving E, van Thiel G, Grobbee DE (2017) Series: Pragmatic trials and real world evidence: Paper 1. Introduction. J Clin Epidemiol 88:7–13. https://doi.org/10.1016/j.jclinepi.2016.12.023
87. Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA (2018) Target Validity and the hierarchy of study designs. Am J Epidemiol 188(2):438–443. https://doi.org/10.1093/aje/kwy228
88. Langan SM, Schmidt SA, Wing K et al (2018) The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ 363:k3532
89. Public Policy Committee ISoP (2016) Guidelines for good pharmacoepidemiology practice (GPP). Pharmacoepidemiol Drug Saf 25(1):2–10
90. Berger ML, Sox H, Willke RJ et al (2017) Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. Pharmacoepidemiol Drug Saf 26(9):1033–1039. https://doi.org/10.1002/pds.4297
91. Pawaskar M, Bonafede M, Johnson B, Fowler R, Lenhart G, Hoogwerf B (2013) Medication utilization patterns among type 2 diabetes patients initiating Exenatide BID or insulin glargine: a retrospective database study. BMC Endocr Disord 13(1):20. https://doi.org/10.1186/1472-6823-13-20
92. Bron M, Guerin A, Latremouille-Viau D et al (2014) Distribution and drivers of costs in type 2 diabetes mellitus treated with oral hypoglycemic agents: a retrospective claims data analysis. J Med Econ 17(9):646–657. https://doi.org/10.3111/13696998.2014.925905
93. Jacobs E, Hoyer A, Brinks R, Kuss O, Rathmann W (2017) Burden of mortality attributable to diagnosed diabetes: a nationwide analysis based on claims data from 65 million people in Germany. Diabetes Care 40(12):1703–1709. https://doi.org/10.2337/dc17-0954
94. Reed M, Huang J, Brand R et al (2013) Implementation of an outpatient electronic health record and emergency department visits, hospitalizations, and office visits among patients with diabetes. JAMA 310(10):1060–1065. https://doi.org/10.1001/jama.2013.276733
95. Reed M, Huang J, Graetz I et al (2012) Outpatient electronic health records and the clinical care and outcomes of patients with diabetes mellitus. Ann Intern Med 157(7):482–489. https://doi.org/10.7326/0003-4819-157-7-201210020-00004
96. WellDoc (2017) WellDoc receives FDA 510(k) clearance to offer a non-prescription version of BlueStar Digital Therapeutic for Type 2 Diabetes. Available from https://www.welldoc.com/news/welldoc-receives-fda-510k-clearance-to-offer-a-non-prescription-version-of-bluestar-digital-therapeutic-for-type-2-diabetes/. Accessed 20 Dec 2019
97. Dennis S, Taggart J, Yu H, Jalaludin B, Harris MF, Liaw ST (2019) Linking observational data from general practice, hospital admissions and diabetes clinic databases: can it be used to predict hospital admission? BMC Health Serv Res 19(1):526. https://doi.org/10.1186/s12913-019-4337-1
98. Williams R, van Staa TP, Gallagher AM, Hammad T, Leufkens HGM, de Vries F (2018) Cancer recording in patients with and without type 2 diabetes in the Clinical Practice Research Datalink primary care data and linked hospital admission data: a cohort study. BMJ Open 8(5):e020827. https://doi.org/10.1136/bmjopen-2017-020827
99. Gordon JP, Evans M, Puelles J, McEwan PC (2015) Factors predictive of weight gain and implications for modeling in type 2 diabetes patients initiating metformin and sulfonylurea combination therapy. Diabetes Ther 6(4):495–507. https://doi.org/10.1007/s13300-015-0134-y