Conducting real-world evidence studies on the clinical outcomes of diabetes treatments

Sebastian Schneeweiss, MD, ScD (0000-0003-2575-467X)
Elisabetta Patorno, MD, DrPH

For: Endocrine Reviews

Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States

Address of Correspondence: Dr. Sebastian Schneeweiss, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, 1620 Tremont Street (Suite 3030), Boston, MA 02120. Phone: (617) 278-0930, Fax: (617) 232-8602, E-mail: schneeweiss@post.harvard.edu

Funding: This work was funded by the Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA. Dr. Patorno is recipient of a career development award from the National Institute on Aging (K08-AG055670).

Acknowledgement: This work is a product of many discussions and published work over the past two decades with colleagues of the Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine of the Brigham and Women’s Hospital. We particularly like to thank Drs. Avorn, Glynn, Gagne, Huybrechts, Rothman, Wang, Walker, who keep encouraging us in doing our best to conduct unbiased studies on drug effects and communicate clearly.

Author Disclosures: Dr. Patorno is investigator of investigator-initiated grants to Brigham and Women’s Hospital from Boehringer Ingelheim, not directly related to the topic of the submitted work; Dr. Schneeweiss is investigator of investigator-initiated grants to Brigham and Women’s Hospital from Bayer, Vertex, and Boehringer Ingelheim unrelated to the topic of this study. He is a consultant to Aetion Inc., a software manufacturer of which he owns equity. These interests were declared, reviewed, and approved by Brigham and Women’s Hospital and Partners HealthCare System in accordance with their institutional compliance policies.

Contributions: Dr. Schneeweiss conceived the article and wrote the first draft. Dr. Patorno contributed her expert knowledge and critically reviewed all aspects of the article.

© The Author(s) 2021. Published by Oxford University Press on behalf of the Endocrine Society.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Abstract

Real-world Evidence (RWE), the understanding of treatment effectiveness in clinical practice generated from longitudinal patient-level data from the routine operation of the healthcare system, is thought to complement evidence on the efficacy of medications from RCTs.

RWE studies follow a structured approach: (1) A design layer decides on the study design, which is driven by the study question and refined by a medically informed target population, patient-informed outcomes, and biologically informed effect windows. Imagining the randomized trial we would ideally perform before designing an RWE study in its likeness reduces bias; the new-user active comparator cohort design has proven useful in many RWE studies of diabetes treatments. (2) A measurement layer transforms the longitudinal patient-level data stream into variables that identify the study population, the pre-exposure patient characteristics, the treatment, and the treatment-emergent outcomes. Working with secondary data increases the measurement complexity compared to primary data collection that we find in most RCTs. (3) An analysis layer focuses on the causal treatment effect estimation. Propensity score analyses have gained in popularity to minimize confounding in healthcare database analyses. Well-understood investigator errors, like immortal time bias, adjustment for causal intermediates, or reverse causation, should be avoided. To increase reproducibility of RWE findings, studies require full implementation transparency.

This article integrates state-of-the-art knowledge on how to conduct and review RWE studies on diabetes treatments to maximize study validity and ultimately increased confidence in RWE-based decision making.

Key words:
Diabetes, real-world evidence, healthcare databases, causal treatment effects, confounding, bias, measurement, regulatory decisions, pharmacoepidemiology
Real World Evidence studies

A transactional database records a stream of healthcare records in calendar time for all enrolled patients:

All patient-level encounters are arranged longitudinally

A causal study design is implemented in event time anchored at the cohort entry date

Validated software ensures transparency and reproducibility of study implementation

Causal effect estimates are derived after propensity score matching and other techniques of balancing treatment groups

© 2021 ENDOCRINE SOCIETY
I. Hopes for and barriers to Real-World Evidence in diabetes care

I.a. Real-world evidence complements RCT evidence: “and” not “versus”

Real World Evidence (RWE), the understanding of causal treatment effects from electronic data generated by the routine operation of the healthcare system, has gained much attention from regulators, payers, and physician groups. RWE is thought to complement essential evidence on the efficacy of medications that we gain from randomized controlled trials (RCTs), by providing information on their effectiveness in clinical practice. Instead of a dichotomy of RCT versus RWE, there is a gradient of increasingly pragmatic aspects in randomized trial designs that gradually approximate RWE studies and vice versa.

The excitement about RWE stems from the limitations of RCTs: They are costly, take time to plan and execute, expose patients to experimentation, and are usually designed to answer a narrow hypothesis in a highly targeted population. Clinicians, however, have an abundance of questions arising from their daily practice that a given RCT may not answer. A specific population may have been excluded from the trial, an outcome not measured, a dose or a combination of drugs not considered, a relevant comparator agent not studied, etc. The hope is that given the variation of care for patients with diabetes in clinical practice, RWE using healthcare databases will address some of those knowledge gaps expeditiously and at low cost without unnecessarily exposing patients to experimentation.

Modern-day healthcare systems produce an enormous amount of longitudinal patient-level electronic data that can be used to correlate medication use with health outcomes in clinical practice. Data sources include (but are not limited to) insurance companies that process claims with coded diagnostic and procedure information, electronic health records with substantial clinical details, patient-reported outcomes from wearable devices that provide information from outside the professional medical system, etc. These data are linked together creating powerful databases that are used by researchers. It is easy to agree that such data should not go unused; rather they should be used to support our learning about drug effects in the very patients we treat.

Recent cardiovascular outcome trials of sodium-glucose cotransporter-2 (SGLT-2) inhibitors to treat type-2 diabetes have demonstrated substantial reduction in hospitalization for heart failure(1-3) and major adverse cardiovascular events.(1,2) These findings have been replicated in non-interventional database studies that make use of ‘real-world’ data to generate ‘real-world’ evidence.(4-7) However, would we have believed the RWE studies in the absence of the findings from randomized controlled trials? Why is it that we have so much more confidence in RCTs than in RWE studies? Skepticism of RWE is justified. There are plenty of examples where RWE studies were in stark contradiction of RCTs. Think of hormone replacement therapy (HRT) in post-menopausal women, for which a postulated reduction in the risk of coronary heart disease was later found to actually increase the risk(8,9); vitamin E supplementation was thought to be protective against coronary heart disease(10,11), but the effects could not be reproduced in a large outcomes trial (12); and the substantial reduction in fractures and dementia associated with statin use in RWE studies were not borne out in RCTs.(13,14)

No matter how evidence is generated, it needs to be internally valid and generalizable to an identifiable target population in order to be actionable. As we care for patients with diabetes,
treatment recommendations must be based on our knowledge of causal treatment effects instead of associations that may be spurious; there are no shortcuts to such medical insights. Our need to understand causal effects, which we usually gain from RCTs, is at the center of this article. We will discuss design, measurement, and analytic strategies that guide investigators and reviewers of RWE to come as close as possible to causal conclusions on treatment effectiveness in patients with diabetes.

I.b. Why is RWE well-suited to diabetes research to complement RCTs?

The suitability of RWE is mostly determined by whether the underlying data sources are appropriate for completing a specific study. Data that can be turned into RWE are those collected through the routine operation of a healthcare system.[15] Most RCTs collect data primarily, which means the researcher is in charge of what to measure, how to measure, and when to measure critical study observations, ensuring high data quality and completeness. Non-interventional studies can also collect primary data, as in well-defined cohort studies where researchers seek to test a specific hypothesis, like the Framingham Heart Study and the Nurses’ Health Study, which have clarified many important aspects of diabetes, its prevention, and its treatment (Figure 1).

Most RWE studies of medical products including antidiabetic medications, however, are conducted with data electronically generated by the healthcare system. This is in essence automatic prospective recording of dispensed and administered medications without relying on patient recall. Every encounter with the professional healthcare system is recorded by insurance claims data and includes diagnostic and procedural information. These are not perfect measurements of health and health outcomes, and such information needs to be recognized as reflecting the recording practices of providers/organizations under economic constraints. This insight is critical in understanding that health care claims data are very useful for some research applications and less useful for others.[16] Many other secondary data sources can be added, including geocoding information, patient sensors, lab test results data, and electronic health records. Registry studies rely on systematic data collection mechanisms that are often investigator designed and operated. They may be based on primary data collection, or can be abstractions from electronic health records, or may contain patient-reported information, or a mix of various data sources. Researchers then embed studies within those longitudinal data collections.

For RWE studies of diabetes, a range of clinical information is critical. Specifically, whether key measurements of exposure, outcomes, and confounding factors are recorded with sufficient precision and completeness in a given real-world data source. Exposure to typical anti-diabetic prescription drugs, either prescribed or dispensed, are well recorded longitudinally and in large numbers of patients. While tablet strength is well recorded, some dosing details may be unreliable or lacking, for example insulin units or dosing by body weight. Key clinical endpoints, for example major cardiovascular events, hypoglycemia, ketoacidosis, etc., are recorded as hospitalizations or emergency room visits with specific diagnostic and procedural codes. Microvascular disease is recorded either by diagnostic codes or by its treatment. HbA1c, fasting glucose levels etc. are recorded well in laboratory test results databases. Confounding factors including duration of diabetes, disease status, and level of disease control can be assessed reasonably well through direct measurement of HbA1c or proxy measurement, for example glucose-lowering treatment patterns, including type and intensity, and diabetes complications.[17] Consequently, RWE studies have been frequent in diabetes care research, although some lack of methodological rigor has been noted.[18-21] The fact that RWE has both successfully replicated and predicted findings of RCTs[4,22] instills confidence that causal conclusions can be reached from RWE in diabetes if implemented correctly.
I.c. Perceived barriers to using RWE for decision making

In addition to obvious criticisms of RWE, like lack of randomized treatment assignment at baseline and the secondary use of data that were not generated for research, which will be addressed throughout this article, a range of additional reservations hinder the use of RWE findings for high-stakes decisions, including drug approval decisions, label expansion, coverage decisions, clinical guideline writing, and prescribing decisions in clinical practice. Surveys show that stakeholders are worried about the complexity of modern-day RWE studies and that they do not understand the methodology and, therefore, cannot interpret the validity of a given study. (23,24)

The reason that decision makers love RCTs is not only the random treatment assignment and study specific primary data collection, but also the logical clarity of typical parallel group RCTs that make them easy to understand and assess their validity. Substantial efforts are under way to improve transparency, pre-specification, and reproducibility of RWE as pre-requisites to allow reviewers to fully assess a RWE study. (25-28) Increasingly sophisticated and integrated RWE software platforms become available that reduce human error, (29) have built-in 100% transparency including audit trails, and allow sharing of study data and analysis with third parties without violating data privacy. (30)

II. RWE and regulatory decision making

II.a. Post-approval safety studies

FDA and other regulatory agencies have a long history of using RWE to assess the safety of regulated medications. For example, RWE played a role in raising the alarm of increased cardiovascular risk with rofecoxib (31) and in alleviating concerns about bleeding risk with dabigatran. (32,33) The Sentinel Initiative, launched in 2008, created a national system of health insurance claims databases that can be used for rapid safety assessment. (34) In general, RWE safety studies such as post-marketing requirements in the US or imposed post-authorization safety studies in the EU can arise as risk management planning at the time of approval, or as part of a rapid regulatory response to a new safety signal that arises at any time after marketing.

Some post-marketing assessment of safety continues to rely on randomized trials. The mandate for studies of cardiovascular safety for newly approved antidiabetic drugs has resulted in more than 20 complete or ongoing cardiovascular outcome trials, although many of them have been initiated with the secondary aim of demonstrating efficacy in preventing cardiovascular events in order to receive supplemental approval for that indication. (35,36)

A key reason that regulators are comfortable with RWE to answer questions about adverse events is that, because most such events are not anticipated by the prescriber, there is little risk-selection at the point of starting the medication, which reduces the risk of confounding bias. (16) Some adverse events are so rare that randomized trials cannot provide conclusive answers, as in the case of diabetic ketoacidosis among users of SGLT-2 inhibitors, where large database studies characterized the signal. (36,37)
II.b. Effectiveness claims based on RWE studies

More recently, regulatory agencies have developed interest in reexamining the value of RWE for substantiating effectiveness claims for approvals, supplemental approvals, or label expansions. The FDA's requirement of “substantial evidence” of effectiveness refers to both the quality and the quantity of the evidence. It provides that all clinical investigations supporting effectiveness should be of appropriate design and of high quality. The quantity of evidence needed may vary in given circumstances, e.g., two adequate and well-controlled studies, one adequate and well-controlled study plus confirmatory evidence, or reliance on a previous finding of effectiveness of an approved drug when scientifically justified. Although randomized superiority trials with placebo- or active-control designs generally provide the strongest evidence of effectiveness, there are circumstances under which studies not using a placebo control or randomization may be acceptable. RWE is thought to fill these gaps.

Figure 2 provides an overview of potential use cases for RWE in regulatory and coverage decision making. As has been done in rare diseases for decades, increasingly highly targeted treatments seek approval with single-arm trials that are matched to external control arms without baseline randomization. The most promising regulatory use case for RWE beyond safety studies in diabetes is secondary or supplemental indications or labeled expansions for agents that are well understood based on RCT evidence. The sponsor wishes to expand the indicated patient population or broaden the labeled benefits, often from a surrogate endpoint to a clinical endpoint. A frequent use case for payer organizations is establishing the comparative effectiveness in clinical practice of multiple alternative treatment options.

Regulators increasingly consider accelerated pathways or conditional approvals acknowledging the evolving nature of the evidence base by requiring RWE studies to better understand the safety and effectiveness of the accelerated-approved drug.

II.c. Regulatory decision making with RWE

Regulatory decision makers understand the complementary nature of RWE and RCT evidence. Among the New Drug Applications received by the FDA in 2019, 60% submitted RWE study findings either as supplemental evidence or as essential evidence. In 10% of those the FDA considered the evidence both sufficient and actionable.

In order for a reviewer to evaluate the validity of a given study he/she needs to first understand exactly how the study was implemented. Particularly because RCTs have a logical causal study design and clear analytic strategy, decision makers report that they can review RCTs and have confidence they understand exactly what was done. They often miss such clear logic in RWE studies buried under analytic complexities. It is thus of great help to think of an ideal target trial to mimic when planning and describing a RWE study (section IVb). That reduces investigator errors and enables reviewers to gain confidence in understanding the study design and being able to review its validity. Complexity in RWE studies is not a strength per se unless it is a necessity. Simplicity in design — for example, a new-user-active comparator cohort design (section IV.b.i) — is often superior to other design variants and easier to explain, analogous to an RCT but without baseline randomization.

Professional societies have agreed on the RWE study parameters that need to be disclosed in detail in order to make a study reproducible and as such fully reviewable. In order to further facilitate the reviewing of RWE studies, a consortium of academics, regulators, and industry
representatives have agreed on a structured reporting template that helps investigators to implement and then report to others exactly how a study was conducted. Another work group determined necessary steps to make the process of conducting RWE studies transparent for regulatory submissions. These activities have evolved into a push to preregister RWE studies on specialized sites (encepp.eu) or generic study registration sites (clinical trials.gov; cos.io) including depositing the structured reporting template or a study protocol. Regulators often wish to see audit trails that enable them to inspect all adjustments to the study design and analysis made since the protocol was registered and deposited, and particularly to determine whether such changes were informed by inspection of preliminary study results. The latter is of course a matter of concern.

Ultimately, regulators are interested in whether a given RWE study lends itself to causal conclusions about the studied treatment effect. While it has been shown in double randomized experiments that RWE studies indeed come to the same finding as RCTs if measurements are accurate and complete demonstrating the validity of modern analytic techniques and causal study designs, the validity of RWE in healthcare still needs to be fully established. Reviews comparing pairs of previously published non-experimental studies with RCT findings often compare studies that analyzed slightly different questions and can establish a lower bound of what agreement to expect. In contrast, ongoing prospectively planned RCT replication projects will establish a more realistic expectation of achievable agreement by reducing emulation differences, and if RWE does not calibrate well against RCTs, we would learn why not.

III. **Real-world data, how they were generated, and their use for research**

### III.a. Longitudinal healthcare data

Modern healthcare systems generate a wealth of electronically stored information on individual patients, producing ongoing data streams that can be connected longitudinally through patient identifiers. Today most pharmacoepidemiology studies use such electronic longitudinal data on medications and health events captured during the routine operation of a healthcare system. These data include insurance claims data, electronic health records, registries, among others. There are several reasons that such data have gained popularity in RWE research. First, they cover populations more representatively than most experimental studies. Second, they include prospective recording of prescribing or prescription-filling in great detail and do not rely on patient consent and patient recall. Third, they do not require experimentation in humans, and are quicker and less expensive than most trials or other studies based on primary data collection. Fourth, the prospective longitudinal recording of healthcare encounters with well-recorded service dates provides clarity on temporality, which is a prerequisite for causal inference of treatment effectiveness.

Healthcare databases are transactional databases that collect clinical and administrative information related to the delivery and administration of healthcare. As encounters occur and services are provided, records are generated and added. Each service provided comes with a date stamp and patient identifier, generating longitudinal patient records of increasing duration (Figure 3A). As a first step in the implementation of a study, one identifies and sets aside a section of the dynamic data stream that covers the calendar time period of interest (Figure 3B). This data snapshot is a prerequisite for making results from a study replicable. It produces an
enumerable set of longitudinal patient records, each with a start and end date in calendar time. Encounters and services are recorded with diagnostic and procedural information on each patient’s timeline (Figure 3C). The rules and algorithms that define a specific study design implementation are applied to each patient’s longitudinal data stream (Figure 3D).

III.a.i. Dates and time windows

Certain principles guide the design and implementation of studies in healthcare data streams. One of the most important is the longitudinality of measurement. Many measurements in healthcare databases are made by reviewing the information recorded during multiple healthcare encounters over time. In contrast, studies with primary data collection establish a study subject’s health state at a point in time when the patient is thoroughly interviewed or examined during a study visit. In healthcare databases, there is no defined interview date with the investigator team, only observation of actual care patterns. Studies rely on the occurrence of visits and other healthcare encounters to collect information that was recorded while providing care. Thus, information that we often conceptualize as being captured at a particular point in time, such as baseline patient characteristics before exposure, is recorded during a defined time window through a series of encounters. (47)

III.a.ii. Anchors in patient event time

When implementing a study of the effectiveness or safety of medications, the time scale shifts from calendar time to patient event time in which specific algorithms define events. As in RCTs, where the randomization date is the most critical anchor date for subsequent analyses, the Cohort Entry Date (CED, also referred to as the index date) is the primary anchor in a non-interventional database study (Figure 3D). The CED is the date when subjects enter the analytic study population.

Secondary temporal anchors are defined relative to the first-order anchor, the CED. Similar to temporal ordering in a randomized trial, we wish to assess all patient characteristics before the start of exposure to avoid adjusting for causal intermediates. (48) Therefore we define an exclusion assessment window and a covariate assessment window (Figure 4). In many applications, we want to make sure that the outcome of interest has not yet occurred at study entry and the beginning of medication exposure. To study such newly occurring events, investigators can require an outcome washout window. Similarly, an exposure washout window of defined duration determines the new use of a drug or other treatment. For example, to identify new users of direct oral anticoagulants, we can require the index dispensing to be preceded by 6 or 12 months during which there was no evidence of anticoagulant use.

The follow-up window, during which the study population is at risk for developing the outcome of interest, begins after study entry. It may begin on the CED, or after an assumed induction window before which there is no biologically plausible effect of exposure on the outcome. The maximum duration of the follow-up window is defined by one or more censoring criteria, e.g., end of enrollment, death, maximum causal time window, or end of data stream. In case-based sampling designs nested in cohorts of patients, like case-control and case-crossover sampling, the study entry can be additionally defined by an event date. Figure 4 exemplifies the longitudinal design choices in a cohort study and a case-control sampling of the same cohort.
III.b. Typical real-world data sources

Much has been written about real-world data sources in healthcare describing in great details the opportunities and limitations of various data types. (16, 49) Here we provide a summary of key aspects.

III.b.i. Claims data

Healthcare claims databases provide a patient-level longitudinal data stream of all encounters with the professional healthcare system, including physician services and hospitalizations, accompanying diagnoses and procedures, and all filled outpatient medication prescriptions, in addition to basic demographic and insurance enrollment information. They contain the billing codes that healthcare providers submit to payers, e.g., private insurances, Medicare, and Medicaid. While they provide a complete longitudinal record of all encounters, unlike electronic health records, which struggle with data leakage from out-of-network services, some databases from insurers with high membership turnover contain limited longitudinal follow-up, making them less suitable for the study of long-term outcomes. Health insurance data vary throughout the world regarding representativeness of the general population, the scope and depth of the included information, data quality and completeness, and linkability with data from other sources, e.g., vital statistics, cancer registries, electronic medical records, and laboratory results. (16) Several national healthcare systems like those in Scandinavian countries have universal life-long healthcare coverage. The use of all health services is documented in national health registries; every person is assigned a unique personal identifier at birth or upon immigration, allowing linkage to a range of clinical registries. (50) Such a special setup ensures immediate generalizability of study findings and allows long-term follow-up studies; given the limited population size and homogenous ethnicity of most Scandinavian countries, may not be able to answer some questions.

The fact that claims data are transaction data collected for administrative rather than research purposes requires researchers to closely examine whether the measurement of key variables is sufficient for a specific study question, i.e., whether the data are fit for purpose. For example, information about clinically relevant parameters like BMI, dietary habits, family history, and certain lifestyle factors like smoking and alcohol use are not reliably recorded in claims or require additional linkage to other data sources like laboratory test results.

III.b.ii. Electronic health records

Much hope has been pinned on electronic health records (EHRs) to supplement the shortcomings of claims data. EHRs are intended for clinical documentation and contain a wide range of patient health-related information, from symptoms, results of physical examinations, laboratory tests and procedures, diagnoses and treatment plans to medical and social history. (51) Data are obtained from multiple clinicians involved in a patient’s care, and are composed of both structured and free text data. EHRs are increasingly used in research as a source for the type of detailed clinical information that tends to be missing from claims databases; for example, Hb\textsubscript{A1c} levels can be extracted from EHR data, and the onset of diabetes is often recorded in physician notes.

A key limitation of most EHR data is that only patient health information generated within the network of the providers that maintain the system will be documented. The resulting “data leakage” yields an incomplete picture if the patient’s care is fragmented and the pertinent information is not electronically shared between providers. Because the original intention of EHRs was not research but support of clinical care, incompleteness is a common challenge to assuring data quality. In addition, there is no universally adopted standard for the types of data that should constitute EHRs,
resulting in heterogeneity between EHR systems. Missing data and between-system heterogeneity in the scope of records need to be accounted for when designing RWE studies based on EHR data.

The emergence of personal health records (PHR) promises to overcome EHR limitations in terms of completeness and heterogeneity. The purpose and components of PHRs closely resemble those of EHRs, the key difference being that patients control and manage access to their own health information with PHRs. Patients can choose to share their health data with providers, for example in cases where patients use home monitoring systems or wearable devices to track the progress of their chronic disease management and activity or document patient-reported outcomes. This flexibility provides physicians with more information to help care for their patients and researchers with more data across the patients’ care continuum.

III.b.iii. Disease and treatment registries

A patient registry uses non-interventional study methods to systematically collect longitudinal information on patients with a particular disease or treatment type from multiple data sources such as EHRs, patient self-report, laboratory results, and surveys. However, because there are no standard definitions, a reviewer must always examine how exactly data were generated and should have reach-through capability to the raw data, as journals like The Lancet now require. Unlike claims and EHR databases, registries may collect very specific clinical information necessary for some RWE studies. Nevertheless, depending on the registry it may not reflect routine practice, and the longitudinal record of medication use is often compiled via patient recall, which carries its own limitations.

A disease registry captures condition-specific information for a cohort of patients, including information on diagnostic testing, prognostic testing, therapies offered and received, family history of disease, behavioral and environmental risk factors, symptoms, and disease progression as well as basic demographic information such as age and gender. The specific data collected as well as the length of patient follow-up – which may range from one episode of care to the entirety of the disease progression – are registry-specific, dependent on the main purpose of the registry. While data granularity is a strength, the size of a registry often does not allow meaningful RWE studies of treatment effects on clinical endpoints.

A treatment registry is in principle similar to a disease registry; however, it includes only patients who receive a specific treatment. Though such registries were frequently mandated by regulators for newly approve medications, it is now widely accepted that their utility is limited, as no information is collected on reasonable comparator patients.

III.b.iv. Vital records

In diabetes research, information on the date and cause of death are important. In the US, these data are available in the National Death Index (NDI) from the Centers for Disease Control and Prevention. They can be linked with claims or EHR databases using patient identifiers. The linkage accuracy is in the high 90 percent range, and the immediate and underlying causes of death are generally considered well recorded if one is interested in larger disease areas, e.g. cardiovascular death, cancer death, or accident, but less accurate for more specific causes.
III.c Turning real-world data into real-world evidence

It is critical to fully understand a data source before attempting to generate evidence on causal treatment effects. The process of planning, implementing, and reviewing a RWE study always comprises three interdependent layers that establish a linear workflow (Figure 5):

(1) A design layer clarifies the basic study design choice, which is best informed by imagining the randomized trial we would ideally perform and want to emulate in real-world data, the target trial. It often guides us toward the new-user active-comparator cohort design (66–68) that has been proven to predict and replicate RCT findings of several diabetes treatments (4,22,59).

(2) The measurement layer transforms the longitudinal patient-level electronic data stream into variables that identify the study population, the pre-exposure health state for confounding control in the absence of baseline randomization, the treatment status, and the treatment-emergent outcomes. Working with secondary data increases the measurement complexity compared to the primary data collection we see in most RCTs.

(3) An analysis layer focuses on the causal treatment effect estimation, considering the data collection mechanism. Propensity score analyses to achieve balance in patient characteristics between treatment groups have gained popularity because of their specific suitability to large secondary databases. Confounding bias and differential follow-up can be further reduced with additional techniques, and well-known biases, like immortal time bias, adjustment for causal intermediates, or reverse causation, can be avoided.

The following chapters will address the three layers.

IV. Study design choice

IV.a. Fundamental consideration in selecting study designs

The clinical study question informs study design choices. In most RWE studies, the design choice is further influenced by the content and limitations of the underlying data sources. When studying the effect of a medication treatment, considerations about the sources for treatment variation, i.e., whether the drug exposure varies within a patient, between patients, or between providers who treat groups of patients, influence fundamental decisions on the appropriate study design, whether it is a non-experimental study or a randomized trial.

In a hypothetical counterfactual experiment, one would treat a patient and observe the occurrence or non-occurrence of the health outcome. Then, counterfactually, one would rewind time, and repeat while leaving the patient untreated, keeping all other factors constant to establish a counterfactual experience. This hypothetical experiment would establish causality in that patient. To approximate this thought experiment, we introduce or observe exposure variation within the same patient at different times, between different patients at the same time, or between providers caring for groups of patients (Figure 6).

If we observe fluctuations of exposure status within a patient over time, e.g., a headache medication, and if that drug has a short hypothesized duration of action, and if the event of interest has a rapid onset (e.g., liver toxicity), then we may consider a case-crossover design or self-controlled case series. Such designs inherently control for time-invariant risk factors by comparing a patient’s experience with himself or herself at different times. While these designs are powerful to study triggers of rapid-onset events, (60) they are less popular in chronic disease
like diabetes as the disease state may worsen with time and with it the treatment may change, which leads to within-person confounding.\(^{(61)}\)

Most RWE studies exploit naturally occurring treatment variation between patients, and therefore use a cohort study design with concurrent controls. Historical control groups have seen a renaissance, mainly because highly targeted treatments reduce the size of the available study population, and investigators resort to observing past experiences to increase the usable data space. Within cohorts, efficient sampling designs like case-control, case-cohort, or two-stage sampling can be used when information gathering is time-consuming or expensive.\(^{(57)}\)

Medication exposure variation between groups of patients through higher-level entities, i.e., between physicians, hospitals, health plans, regions, etc., can be exploited using instrumental variable analyses. The instrument, which corresponds to the reason for the group-level exposure variation, must be unrelated (directly or indirectly) to patient characteristics.\(^{(62)}\)

Selecting a comparator group is a fundamental design choice and substantially influences clinical interpretation and may strongly alter the effect size. The comparator needs to be relevant in the clinical context and a viable alternative to the study drug. Ideally, we want to restrict the comparison population to patients who, in clinical practice, have the same indication as the users of the study agent. The oral antidiabetic drugs rosiglitazone and pioglitazone are an example of such a medication pair. They were marketed around the same time, were both indicated for second-line treatment of diabetes, come from the same class of compound, and in the early marketing phase were thought to have similar effectiveness and safety profiles. This should make treatment choice largely random with regard to patient characteristics and should therefore make treatment groups comparable by design, resulting in little confounding.\(^{(63)}\)

IV.b. Cohort studies and target trial thinking to avoid bias

There are many ways in which RWE studies are different from RCTs. RWE studies aim to include a wider range of patients and are embedded in healthcare delivery systems, reflecting clinical care as part of routine operation. However, RWE and RCTs are more similar than different, as both try to establish causal relationships between medical products or interventions and health outcomes. Before intervening on patients, we want to ensure that we are treating with medical products that improve health outcomes.

It has therefore been proposed multiple times over decades, and most specifically by Hernan and Robins,\(^{(64)}\) that by envisioning a target randomized trial that one would wish to conduct if it were logistically and ethically possible, and emulating that target trial in the design of a RWE study, then, even in the absence of baseline randomization avoidable design biases will be reduced and clarity about study design increased. Thinking about emulating a target trial encourages clarity in the temporality of when patient characteristics, exposure, and outcomes are measured relative to study entry, which is critical to enable causal conclusions. It clarifies the analytic strategy of an ‘as-started’ (aka intention-to-treat analysis) or an ‘as-treated’ analysis. Once a target trial is conceptualized, the design of the trial-emulating RWE study and potential diversion from the trial reveals potential weaknesses in data quality, data completeness, and in causal inference. It is hoped that such clarity will lead to adjustments in the RWE study that improve validity.\(^{(65)}\) However, trial-emulating RWE study design often exposes a tension between the objective of highly generalizable findings and necessary restrictions imposed to ensure high validity of findings and causal conclusions.
Target trial conceptualization helps identify and avoid design flaws like immortal time bias, adjustment for causal intermediates, reverse causation, and dealing with time-varying hazards and depletion of susceptibles. In a cohort study setting, target trial conceptualization guides users to the new-user study design, as that is what occurs in a typical randomized clinical trial. Studying ongoing users who have “survived” on the drug of interest until cohort entry is known to produce misleading findings.

IV.b.i New user cohort

The most frequent study design by far to evaluate the effectiveness of drugs in diabetes is a parallel group randomized trial. Grossly simplified, patients of interest are identified (e.g., patients with type-2-DM and within a certain range of HbA1c value), there is a washout period during which patients reduce their current drug use to what the trial prescribes, and after that patients are randomized to a treatment group and followed for the occurrence of an event (Figure 7). An emulation of this trial with real-world data would similarly focus on new users and require some non-use period to mimic washout, before patients start treatment (Figure 7). Of course, in a RWE study randomization (R) is replaced with selection (S), which needs attention in the statistical analysis.

There are several advantages to studying new users of a study drug in RWE studies, particularly when comparing them to new users of a viable alternative treatment. As patients in both treatment groups are newly started on medications, they have been evaluated by a physician who concluded they would benefit from starting or escalating therapy with a newly prescribed drug. This process produces comparable treatment groups that are similar with respect to characteristics that are both observable and unobservable in a given data source. The clear temporal sequence of measuring confounders before treatment begins avoids the mistake of adjusting for the consequences of treatment, i.e. causal intermediates. Because of the well-defined starting point of new-user cohorts, it is possible to assess how hazards vary with duration of treatment. Because the new user cohort study design closely emulates the standard parallel-group randomized trial, it is more easily understood by general readers. This cogency should not be underestimated in an era where decision-makers ignore non-interventional studies because they seem too complicated and the validity of the study seems obscure. Examples of such new user cohort studies are the risk of psychosis in children and adolescents starting stimulant medications or the effects of statin use on a range of health outcomes.

The epidemiologic studies on hormone replacement therapy (HRT) and coronary heart disease (CHD) exemplify the issues arising when not matching a precise study question with the right study design. Such studies showed a strong association between ongoing HRT use and a reduced risk of CHD after extensive confounding adjustment. This result was famously refuted by a subsequent randomized trial, the Women’s Health Initiative, showing cardiovascular harm. But a landmark re-analysis of the Nurses’ Health Study data conformed with the RCT result. The authors did this by structuring the observational analysis to mirror the clinical question asked in the trial, i.e. how does starting HRT in a post-menopausal woman affect her cardiovascular risk? The earlier observational studies studied ongoing users of HRT which implied effectively excluded women with CHD events that occurred shortly after initiation of HRT, a remediable design bias known in pharmacoepidemiology as “depletion of susceptibles” or in other fields as “survivorship bias.” No amount of adjustment for baseline factors in elaborate regression analyses between ongoing users and non-users can overcome that, but refining the question, and tuning the design and analysis to the question can.
The new user design also has clear benefits when studying newly marketed medications: it avoids comparing populations composed of first-time users of a newly marketed drug with a population comprising mainly prevalent users of an existing comparator drug.(58) Such a comparison would be prone to bias, because patients who stay on treatment tend to be those who tolerate it, experience its benefits, and may be less susceptible to the event of interest. The same phenomenon plays out when comparing patients who switch medications versus those who stay on their medication. This is an essential clinical question in most chronic conditions, but because patients who experience treatment failure are more likely to switch to an alternative treatment, confounding may be substantial. In the setting of studying newly marketed medication to treat chronic conditions, most patients starting on the new medication have already been using an alternative medication, often one of the comparator drugs. Given the temporal sequence -- that patients first were new users of the comparison drug and subsequently switch treatment to become a new user of the study drug -- the standard new-user design that excludes drug switchers would yield a smaller cohort of study drug users. To accommodate this setting, an alternative to the new user design is the prevalent new user design, which allows study drug users to have previously used the comparison agent.(71) Although this design increases the number of study drug users, which is frequently the factor limiting study size, it increases the risk of confounding by allowing drug switchers to be included in the analysis.

IV.b.ii  Active comparator

Non-user comparisons that are conducted in an attempt to emulate placebo-controlled trials often suffer from strong treatment selection. Persons prescribed the drug differ from those not prescribed the drug in ways that are difficult to completely measure and control analytically. Such strong confounding also occurs when comparing two different treatment modalities, e.g., oral antidiabetic medications versus injectable insulin, or medical treatment versus dietary control. An extreme example of such uncontrollable confounding is the comparison of medication treatment versus implantable cardioverter-defibrillators in patients with heart failure and the risk of sudden cardiac death.(72) The frailest patients will not undergo surgery because of its risks, and yet these patients are at the highest risk for the outcome of interest, biasing the comparison. An example of a successful new-user active-comparator design was the RWE study predicting the findings of the CAROLINA trial months before it was released (Figure 8). New users of linagliptin were compared to new users of glimepiride among patients with type-2-diabetes, mimicking the RCT exclusion criteria as best as possible. Like the trial, the RWE predicted no difference in the cardiovascular composite endpoint but a substantial clinical benefit in reducing the number of hypoglycemic events that lead to emergency room visits or hospitalizations.(22)

IV.b.iii  Immortal time bias

Comparing new users in two active treatment groups not only leads to more comparable patient groups, but also further reduces the chance of immortal time bias, a problem that emerges if future information is used to define earlier exposure status in healthcare databases. A typical example of immortal time bias is to define the group of non-users as patients who do not use the study medication during follow-up for, say, 12 months. By definition, these non-user patients cannot die during the first 12 months of follow-up, or else they could not be included. As their mortality rate must be zero during the first 12 months of follow-up, the inclusion of those 12 months (needed to establish the exposure definition) will bias findings when studying mortality or endpoints with a high risk of dying.(73)
Another example of immortal time bias is found in recent studies of SGLT-2 inhibitors and their mortality benefit. Mimicking a randomized trial, the investigators focused on new users of SGLT-2is and compared them against starters of other glucose-lowering drugs (oGLD) (Figure 9). These patients were then followed for the occurrence of death. So far a straight-forward design. In order to enrich the study, investigators opted to re-categorize people who started on an oGLD and then switched to an SGLT-2 as new SGLT-2 users. Though this would not be problematic in itself, the investigators chose to disregard the prior exposure to an oGLD but still used the accrued person time (see the second hypothetical patient in Figure 9). That meant, as Suissa pointed out, (21) that if patients died on an oGLD they never had the chance to move on to an SGLT2 and were never categorized as such; in turn all oGLD users who switched to SGLT-2s did by definition survive their oGLD use. Attributing the latter immortal person-time to the SGLT-2 experience makes SGLTs appear falsely superior regarding mortality.

IV.c. Cohort sampling designs

Several sampling designs within study cohorts are applied to increase the statistical efficiency of an effectiveness analysis (see Figure 6). Such sampling designs are often relevant when additional information unavailable in the study data source needs to be acquired through great effort or by expending substantial resources. (76)

IV.c.i. Case-control sampling

The odds ratio computed from a case-control study using a sampled risk-set or a random sample of person-moments estimates the rate ratio of the underlying cohort study (see Figure 4). (77,78) An inherent property of the case-control sampling design is that one will obtain the same rate ratio estimate as that from an analysis of the underlying cohort that gave rise to the cases and controls, as also illustrated in empirical examples. (79) When conducting case-control sampling in healthcare database studies the underlying cohort that gave rise to cases and controls is identifiable and enumerable. This is not the case in many community-based or hospital-based case-control studies, where the true underlying source population remains unknown.

There are several good reasons, even in database studies working with previously collected information, to apply case-control sampling. (80) First, it is at times necessary to collect data on a confounder not available in the database by reviewing source data, like duration of diabetes, diet control, BMI, family history, etc. This may be costly and time consuming, so that the greater efficiency of case-control sampling is welcome in such settings. (81) Second, there are drug safety surveillance programs that focus on specific endpoints, particularly if such endpoints require adjudication or special expertise in their assessment and classification, for example a system focused on severe liver injury due to drug exposure. Such systems have established an elaborate system to identify and validate the outcomes of interest and screen a wide range of medications for the incidence of liver injury. (82) Finally, a subtler point of case-control sampling designs is that they make it convenient to study the triggers of an acute event by flexibly modeling the exposure window at varying proximities to the event of interest. (83,84)

A well-described limitation of case-control designs is the inability to directly estimate incidence rates. While that is still the case, in database studies we are usually able to enumerate the underlying cohort in such data and therefore can determine the sampling fractions of cases (often 100%) and of controls, and indirectly establish incidence rates. (85)

IV.c.ii. Case-cohort and 2-stage sampling

Biomarkers observed before cohort entry are often used to stratify a study population of patients with diabetes, e.g. HbA1c values, genetic markers, or retina scan readouts. Here again, such
information is costly to retrieve and therefore a variation of case-control sampling, namely case-cohort sampling, is applied.\(80\) In contrast to simultaneously identifying cases of the study endpoint and sampling controls, in a case-cohort analysis investigators identify a random subset of patients upon entering the underlying cohort and collect the biomarker information. These controls can be repeatedly reused during follow-up.\(86,87\) The design improves upfront planning of biomarker collection, e.g., if frozen specimens need to be thawed, and unlike a case-control sampling, it enables studying multiple endpoints simultaneously like a full cohort study.\(80\)

The two-stage sampling is a combination of both with known sampling fractions at cohort entry and at the study endpoint, thus further improving study efficiency.\(88\) The planning and analysis is more complex and requires the implementation of an elaborate sampling system.\(89\)

V. Measurement considerations when working with secondary healthcare data

Much has been written about data standardization and how to improve data quality. In the end, all discussions of data quality culminate in the same question: Are the data appropriate for this specific study question? It has been shown in a double randomized experiment that if measurements are sufficient, then non-randomized studies can produce unbiased estimates like an RCT.\(43\) Even in a narrow therapeutic area like diabetes research, no single data source or standardization method can answer all questions. It comes down to how the exposures, outcomes, and confounding factors, as listed in Figure 5, are measured. We know which measurement characteristics we should look for; however, they are almost never directly observed in real-world data, and there is little agreement on what measurement performance is sufficient (Table 1). RCTs deal with similar issues and spend much time and money on improving measurements. It may be impossible to certify entire databases as fit for RWE, but we can illuminate the process of data generation and curating up to the point when the data are used for a specific analysis. This allows for an assessment of measurement characteristics, which then informs quantitative bias analyses.\(90\)

The remarks in this section are largely generic or may apply to frequently encountered situations in the U.S. As measurement issues vary highly between specific data sources this section is meant to illustrate principles but cannot replace a detailed understanding of the study data source. This often involves many conversations with the chain of people involved in generating the data, e.g. physicians, coders, insurance companies, or data vendors. It often takes months to fully understand how all information in a given database is to be interpreted.

V.a. Identifying the study population

Clear identification of the study population is of importance to assess the generalizability of findings. In RWE on the treatment of diabetes, most cohort inclusions start with the treatment of interest or a comparator. Exclusions follow based on the desired age range, the absence or presence of certain diagnoses, and the presence of markers of the status of the diabetes and its control. Typical markers of interest are an Hb\textsubscript{A1c}, measurement, duration of diabetes, and BMI, to name three recurrent themes. Now, if those markers were really vital to the interpretation of findings, one would identify a data source that captures these parameters. However, often data sources large enough to support meaningful inferences regarding treatment effects will not contain those measurements. The investigator must then decide whether the fact that some patients got started on certain treatments
is sufficient to correctly categorize the patient population according to their diabetes severity or whether one should spend time and resources to collect the missing information. As is so often the case in RWE studies, investigators face the challenge of deciding between two suboptimal choices, and yet one may be more suitable for the purpose of a specific study.

V.b. Treatment exposure

For RWE studies it is fundamental to measure the start and end of a treatment of interest. Electronic pharmacy dispensing records are considered largely accurate in recording the start of a drug exposure, because pharmacists fill prescriptions with little room for interpretation and are reimbursed by insurers based on detailed, complete, and accurate electronically submitted claims.

Drug dispensing claims contain a field for the number of days supply, representing the number of days a dispensing is intended to cover. Alternatively, one can multiply the number of dispensed pills by their strength and divide by a typical dosing like that described by defined daily doses (DDD) or other estimates. To assess the overall duration of medication use, individual dispensings need to be strung together. While this is a reasonable approach, it may cause exposure misclassification in two ways. If the calculated or pharmacist-recorded days supply is too short or if patients decide to stretch a prescription by using a lower dose, e.g., by tablet splitting, some person-time will be classified as unexposed when it is actually exposed. Most drugs treating chronic conditions are used for extended periods, necessitating multiple refills. A patient can thus be classified as being unexposed intermittently despite continuous exposure. Many investigators, therefore, extend the calculated days supply by some fraction called a grace period, e.g., 10 days per prescription, to avoid this misclassification. However, this strategy can also lead to unexposed time being classified as exposed if a patient discontinues drug use without finishing the supply. The right balance between improved sensitivity versus specificity of drug exposure assessment depends on how well the days supply is calculated, which in turn depends on the type of drug, the underlying disease, and eventually how regularly it is taken.(16)

The addition of a grace period is intended to improve the measure of patient time on treatment. This issue is conceptually different from the exposure risk window, the time period during which an event might be causally attributed to the study drugs. The latter is defined by pharmacokinetic parameters and the underlying biology of the condition.

Electronic dispensing data in most systems accurately record the strength in mg per unit of the dispensed drug in mg per unit. Combined with prescribing information on the frequency of tablets per day dose-stratified effects can be studied. However, some dosing is weight-dependent and requires recently recorded body weight.

In summary, there rarely is a perfect algorithm to classify longitudinal medication exposure 100% accurately in healthcare databases, although data quality is still considered better than self-report and physician notes.(91,92) The choice of measurement strategy depends not only on whether one needs to be more concerned about falsely classifying person-time as exposed or unexposed, but also on the pharmacology of the hypothesized drug effect.

V.c. Diabetes-relevant outcomes

Because utilization databases often lack detailed clinical information, researchers must consider the possible effect of outcome misclassification. Generally, a lack of specificity of the outcome measurement is worse than a lack of sensitivity. A relative risk estimate is unbiased by outcome misclassification if the specificity of the outcome assessment is 100%, even if the sensitivity is substantially less than 100%, as long as the misclassification is non-differential.(93) Studies on the
misclassification of claims data diagnoses, using medical records review as the gold standard, revealed that the sensitivity of claims diagnoses is often less than moderate, while their specificity is often high. This pattern arises because if a diagnosis was recorded, coded, and submitted, it is highly likely that this diagnosis was actually made, particularly in hospital discharge summaries.

Relevant cardiovascular outcomes that lead to hospitalizations, like myocardial infarction, coronary revascularization, acute coronary syndrome, cardiac death, heart failure, and stroke are well captured. Other, microvascular endpoints or adverse events like hypoglycemia, diabetic ketoacidosis etc. often lead to emergency department visits followed by procedures or medications specific to the event. While all these events are well recorded the definitions miss less serious events that are managed successfully by a primary care physician. Those events can be identified by lab test results during follow-up; however, in practice such efforts struggle with non-uniform timing of lab tests in clinical practice that are subject to scheduling and emergency encounters. Many systems that capture lab test results are confined to a specific provider network and once a patient seeks care by a physician outside the network such test results are not available.

Diagnoses for ambulatory services may include diagnostic codes for services to rule out a condition, e.g., a blood glucose test to rule out diabetes. Requiring two or more recordings of a diagnosis made in an ambulatory setting, possibly combined with a procedure claim specific to a confirmed diagnosis, can help to increase specificity.

V.d. Confounding factors

Baseline confounding factors will be measured before the treatment of interest starts to avoid adjusting for causal intermediates. Some challenges in healthcare databases are the complete and accurate measurements of important predictor variables like duration of diabetes, BMI, HbA1c, LDL cholesterol level, creatinine level, etc. Misclassified or unobserved confounder information leads to residual confounding, which is addressed in depth in the section on data analysis.

V.e. Missing data

Missing data is an issue that cuts across all aspects of measurement discussed in the previous sections and is often a reason why specific studies cannot be conducted in a given data source. If critical data items were not recorded or recorded with substantial missingness or misclassification then the data will not be fit for purpose. It is critical to accept that some studies cannot be done with RWD and that other data sources or primary data collection may be necessary to achieve the required measurement characteristics. The question in specific settings is often whether a precise measurement is required for valid inference or whether a proxy measurement or some value imputation is adequate.

Generally speaking, claims data that contain information on diagnoses and procedures lead to misclassification of information not missing data. We usually equate the absence of a code with the absence of the disease. If this is not true then the variable is misclassified but it has a value. Validation studies such as those quoted above will quantify the amount of misclassification and on can make an assessment of whether a study can still be validly performed and effect estimates can be corrected modeling the potential bias.

In electronic medical records, however, we observe many test results values etc. that cannot simply be set to a fixed value. The fact that a certain test was not ordered by the treating physician is in
itself much information and the resulting strategy of including a missing value indicator in the analysis has been shown to be useful except for extreme circumstances.(106,107) Other ad hoc strategies include imputing the mean of all observed values or carrying the last observed value forward, neither of which are satisfying but are often applied.(108)

Missing data issues are addressed in different ways depending on their use. To identify the relevant study population, one prefers a population definition that assures that only patients of interest are included, for example patients with Type-2 diabetes and not type-1 diabetes. With many database studies one has the luxury of large numbers and may want to exclude patients where the recorded data provide no certainty of the disease state.

Missing data in pre-exposure patient characteristics that may be used for confounding adjustment, for example the percent HbA1c or duration of diabetes, may be missing for some patients in a medical records database and in most claims databases are not recorded at all. It can be assumed that the missingness of HbA1c values is not random as the ordering of the test or recording of the disease duration depends on the disease state, on the frequency of office visits, on the treatment, and the progression of the underlying diabetes.(108) The new-user active comparator design, high-dimensional proxy adjustment, and multiple imputation are strategies to mitigate residual confounding caused by such risk factors.(57,109,110) This is addressed in the section on data analysis.

VI. Data Analysis

There is a range of causal parameters that can be estimated in longitudinal studies. Here we focus on those most relevant in pharmacoepidemiology. Deciding on which approach to use is inevitably a tradeoff between the clinical relevance of the different target parameters for the given question and feasibility in being able to obtain unbiased estimates of the targeted parameter.

VI.a. Causal effect of interest

VI.a.i. The as-treated effect

The as-treated effect is the effect of initiating the study treatment and continuing to receive it. Patients’ follow-up time is censored at discontinuation of the initial treatment. The numerical value of the as-treated effect from a given study takes into account the duration of treatment persistence.

The as-treated effect is in most situations of great interest to patients and physicians, because it is informative about the expected treatment effect while the patient is actually using the medication. As-treated analyses need to take informative censoring into account, as discontinuation of the initial treatment may be informed by early signs of the outcome of interest or be otherwise associated with the study endpoint. In many situations, treatment discontinuation is a quasi-random process, and predictors of discontinuation are not obvious. In other situations, discontinuation and switching to another treatment or dose is part of a clinical strategy.

VI.a.ii. The effect of complex treatment strategies

In many chronic conditions, it is recommended to start, stop, switch therapy, or change dose depending on clinical markers. One may therefore be interested in estimating the effect of a treatment strategy instead of analyzing the effect of treatment with a single drug. The as-treated effect in such situations is defined as following a specific strategy, and censoring occurs if a patient
deviates from that strategy. These analyses can become complex and require dynamic control for confounding beyond baseline factors, because future treatment choices depend on clinical markers and health states that change over time. Such time-varying confounding can be addressed with time-varying propensity score weighting approaches embedded in marginal structural models or G-estimation. In diabetes care, consider studying the effect of the clinical strategy to treat to an HbA1c target between 7 and 8% versus not.(111) Treatment choice therefore will vary over time as a function of HbA1c, a marker of disease activity.(112,113) Other examples are the treatment of HIV infection with antiretroviral therapy that may be intensified based on CD4 counts or viral load, which are blood markers for disease activity.(114) Data sources for such studies need to contain frequent measurements of the clinical markers determining treatment choice over time, something that we rarely observe in primary care databases.

VI.a.iii. The as-started effect

The as-started effect is the effect of the initial treatment choice, regardless of whether that treatment continued over a given period of time. It is similar to the intention-to-treat effect in RCTs. The magnitude of the as-started effect from a given pharmacoepidemiology study depends on the specific patterns of deviation from the initial treatment choice during the follow-up time. As patients discontinue treatment, their exposure status will still be categorized according to the initial treatment. This handling avoids issues arising from informative censoring but leads to exposure-misclassified person-time. Two studies of the same treatment strategy, conducted in the same population, could have as-started effects different from the as-treated effect if persistence patterns differ between the studies, but would have the same as-treated effect in the absence of time-varying hazards. If discontinuation is non-differential between compared groups, it will lead to increasingly smaller effect sizes with decreasing as-treated durations and resulting increasing exposure mischaracterization. The more person-time that is mischaracterized, the more the as-started effect estimate will differ from the as-treated estimate. Given that persistence patterns are often less than optimal in clinical practice even for life-extending medications(115), this mischaracterization can be a major concern, particularly if the intended estimate of interest is really the as-started effect.

The as-treated and as-started effects are the same for one-time exposures, like some vaccines, medical device implementation, or very long-lasting drugs like some osteoporosis medications administered once a year. In these settings, adherence beyond the first exposure is much less or not at all relevant. Depending on the medication of interest and the biologic disease model, the as-started effect may be more informative than an as-treated effect. For example, there was a concern that TNF-alpha blocking agents in the treatment of rheumatic diseases could be associated with increased risk of B-cell lymphoma. The postulated disease model was that even fairly short-term exposure to the agents could induce damage, and it would take years for the lymphoma to become apparent. An as-treated effect estimate would not fit this disease model well.(116)

In summary, the as-treated effect is often the clinically most relevant effect, while the as-started analysis requires the fewest assumptions to obtain an unbiased causal estimate.

VI.b. Channeling of treatment to patients and confounding of causal treatment effects

Physicians prescribe drugs in light of disease severity and prognostic information available at the time of prescribing, like duration of diabetes, HbA1c status, microvascular status, prior diabetes complications, renal function, cardiovascular comorbidity etc. The factors influencing this decision vary by physician and with time and frequently involve clinical, functional, or behavioral characteristics of patients that may not be completely recorded in healthcare databases. If such
prognostic factors are not balanced between drug users and comparator patients, then failing to control for such factors may lead to confounding bias. The same mechanism applies to treatments other than drugs, like surgical interventions and medical devices. Because treatment selection according to disease severity and prognosis is an integral part of practicing medicine, the resulting bias can be strong. The confounding arising through selective treatment decisions in medicine is sometimes more specifically called confounding by indication, confounding by contraindication, channeling, healthy-user bias, or sick-stopper bias, all of which address the same underlying challenge.

A first step in resolving potential confounding is understanding the channeling process. This is done by consulting physicians in order to understand the factors practitioners consider when making treatment decisions for a specific disease state, by reviewing treatment guidelines, and by empirically describing the actual prescribing behavior in the study population. The latter often results in tables listing observable pre-exposure patient characteristics by the treatment groups of interest. If meaningful imbalances in key prognostic factors raise concerns, investigators may limit the study to more homogeneous patient subgroups or consider alternative comparison groups. Both strategies may result in more balanced patient characteristics, indicative of equipoise in the clinical decision. If extreme imbalances persist, one should reconsider the feasibility of the study with non-interventional designs.

IV.c. Analyzing comparable patients

VI.c.i. Restriction to similar patients

Restriction is a common and effective analytic tool to make treatment groups more comparable and therefore to reduce residual confounding. Some restrictions are obvious since they are made by explicit criteria, such as for example, limiting the study population to patients 65 years or older and with dementia to study the safety of antipsychotic medications used to control behavioral disturbances in this population. Other restrictions, like matching on a confounder summary score, either a propensity score or a disease risk score, are frequently used in pharmacoepidemiology. It is important to understand the specific reasons for restrictions to reduce bias and their implications for the generalizability of findings.\(117\)

Restricting the study population to new users of the study agent or a comparator agent implicitly requires that both groups have recently been evaluated by a physician. Based on this evaluation, the physician has decided that the indicating condition has reached a state where a treatment should be initiated. Therefore, such patients are likely to be more similar in observable and unobservable characteristics than nonusers or ongoing users of another treatment.

Matching by a summary confounder score, like an exposure propensity score or a disease risk score, has become commonplace in pharmacoepidemiology studies that use secondary healthcare data. Most matching strategies produce a group of exposed patients that have been paired with comparator patients who are similar with regard to the matching variable or score and a remainder group for whom there were no good matches in the comparison group. The latter group, lacking comparators, will not contribute to the analysis.

While restriction is an important tool to improve internal validity, it will reduce generalizability of findings to the patient groups excluded from the analysis. Given that pharmacoepidemiology studies inform decisions that will affect patient care, we place high value on internal validity, even if that comes at the price of reduced external validity. Investigators need to be aware of this trade-off and justify their choices accordingly.
VI.c.ii. Propensity score analyses

Propensity scores (PS) are multivariable balancing tools that can efficiently balance large numbers of covariates, even if the study outcome is rare, which is frequently the case in pharmacoepidemiology. PS analyses have therefore emerged as a convenient and effective means for adjusting large numbers of potential confounders in pharmacoepidemiology studies. They fit the target trial paradigm (alluded to earlier), as the PS emulates the randomization process based on observed data. In a new user cohort design, a PS is the estimated probability of starting treatment A versus starting treatment B, conditional on all observed pretreatment patient characteristics. Estimating the PS using logistic regression is uncomplicated, and strategies for variable selection are well described.(118) Once a PS is estimated based on observed covariates, there are several options to use it in a second step to reduce confounding. Typical strategies include adjustment for quintiles or deciles of the score with or without trimming, matching, fine-stratification, or weighting by PS.

Matching on PS in a cohort study has several advantages that may outweigh its drawback of not using the full dataset in situations where not all eligible patients match. Matching excludes patients in the extreme PS ranges, where there is little clinical ambivalence in treatment choice. Dropping such patients from the analysis reduces residual confounding and may lead to more clinically relevant findings.(119) In contrast to traditional outcome models, PS-matched analyses, particularly fixed-ratio matching, allow the investigator to demonstrate the covariate balance achieved in the final study sample. Table 2 serves as an example of select pre-exposure patient characteristics from a new-user active-comparator cohort study of linagliptin and glimepiride.(22) It shows that while some pre-matching imbalances exist (in green highlight) these differences are balanced after 1:1 propensity score matching with a caliper of 0.05. Post-matching c-statistics or standardized differences of covariates have gained popularity in PS matching analyses.(120) Fixed-ratio matching in cohort studies, such as the frequently used 1:1 matching on PS, does not require matched analyses to obtain an unbiased result. In settings with very few events, 1:many matching or fine stratification may be preferred.(121)

VI.c.iii. Reducing unmeasured confounding: High-dimensional proxy adjustment

Any pre-exposure patient information recorded can be considered potential confounding factors. If the optimal measurement of these factors is not in the investigator’s control, confounding from unobserved factors can be reduced by measuring and adjusting observable proxies of the underlying confounders. To the extent proxy measurements are correlated with the underlying confounders, the unobserved confounders are adjusted.(122,123) Examples of well-measured proxies are the use of oxygen canisters (correlated with frail health), regular use of preventative services (correlated with health-seeking behavior), or glucose-lowering medication use (correlated with HbA1c measurements) etc.

Proxies can be efficiently generated by turning codes/features that were recorded before the start of the exposure into variables like inpatient diagnoses, outpatient diagnoses, procedures, word stems from free text notes etc. For each such generated variable, additional attributes can be assigned, including how frequently the code is recorded and the time elapsed between the code and the initiation of treatment.(124) This results in high-dimensional covariate spaces with several thousand covariates, some of which are true confounders. Principled variable reduction techniques reduce these thousands of covariates that may be confounders to few hundred that are highly likely confounders before entering them into a propensity score model.(109,125,126)

The resulting high-dimensional propensity score is often superior in terms of bias reduction across a range of research questions and versatile in a variety of data sources and coding systems, like the FDA Sentinel analysis showing an increased hypoglycemic event risk among users of the oral anti-diabetic drugs glyburide versus glipizide as observed in randomized trials.(127,128) Its properties are
well understood based on empirical studies and statistical simulation experiments. In diabetes research such data-adaptive proxy adjustment has resulted in well-balanced characteristics that were unobserved in claims data, like duration of diabetes, BMI, HbA1c, LDL cholesterol level, creatinine level, but could be verified in medical records.

VI.c.iv. Reducing unmeasured confounding: Instrumental variable analyses

An instrumental variable (IV) is an observed variable that causes, or is a marker of, variation in the treatment that is unrelated to any risk factors for the outcome and therefore correspondent to random treatment choice. The prototype IV is the coin toss in a randomized experiment: It is closely, often perfectly, related to the exposure status and it is unrelated to patient factors and the outcome.

In non-interventional research, identifying valid instruments is difficult and valid IV analyses of drug effects are infrequent. An example of an IV is a hospital drug formulary, a preference-based IV. Some hospitals use only drug A for a given indication, and other hospitals that are comparable in their patient case mix use only drug B. It is a reasonable assumption that patients do not choose their hospital based on its formulary preference but instead on location and recommendation. Therefore, the choice of drug A versus drug B should be independent of patient characteristics in the hospitals with these restricted formularies. If no disease state-related factors lead to preferential admission to one of the hospitals, comparing patient outcomes from drug A hospitals with patient outcomes from drug B hospitals should result in an unbiased estimate of the effects of drug A versus drug B.

In principle, treatment preference can be influenced by time if treatment guidelines change rapidly and substantially as observed with the introduction of DPP-4 inhibitors as second-line treatment in the US. A comparison of patient outcome rates before versus after a sudden change in treatment patterns may then be a reasonable approach. More commonly, IV analyses utilize individual, clinic/hospital, or regional treatment preferences. For example, one may use physician-prescribing preference to study the effect of analgesic treatment with COX-2 selective inhibitors versus non-selective NSAIDs on the risk of upper gastrointestinal bleed. A study demonstrated that such preference is a fairly strong and valid instrument. Others used regional variation in the rate of cardiac catheterization to estimate its effect on mortality after myocardial infarction. While this regional preference instrument was weaker than the physician prescribing preference, it was argued that the instrument was more valid, as it is less likely that patients would move to another region to receive the preferred care, but they may more readily switch their physician.

The price of potentially unbiased estimation in IV analyses is the ultimately untestable assumptions that the authors need to argue based on substantive knowledge and some empirical data. Because of the two-stage estimation, IV analyses are generally less precise, which may limit their utility for decision-making. Practical guidance exists for conducting IV analysis in RWE.

VI.c.v. Reducing unmeasured confounding: Utilizing data-rich sub-samples for improving inference

If unmeasured patient characteristics that are deemed important confounders remain unobserved, then additional information can be collected in a subset of patients. A common version thereof is case-control sampling or case-cohort sampling, where only a sample of controls or a sample of exposed and unexposed subjects will be used to collect detailed confounder information (see section above).

It is increasingly possible to link information-rich electronic health records or registry data to large subsets of population-based claims data studies. Such linkage can be used to demonstrate the balance achieved in patient characteristics that were unobserved in the claims data of the main study. In a new-user active-comparator cohort study comparing linagliptin versus pioglitazone...
regarding cardiovascular outcomes investigators used 1:1 PS matching on more than 100 pre-exposure baseline characteristics that were observable in claims data to achieve comparable treatment groups.(17) In the linked EHR data it was then demonstrated that laboratory test results, BMI, and duration of diabetes were well balanced between treatment groups, although these parameters were not part of the claims data analysis and only observable in the subset of EHR-linked patients (Figure 10).(17) In the right data environment, such data linkage is possible, and one can routinely check the achieved balance of baseline factors unmeasured in the main study.

There are multiple strategies to incorporate additional detailed information on confounding factors that is available in a subset of subjects into the main study and adjust the initial observed effect estimate for any residual confounding. Simple algebraic solutions are available to adjust for individual binary factors, as was demonstrated in a study of older adults and missing quantifications of the limitations in activities of daily living, cognitive impairment, and physical impairments.(105) PS calibration extends this simple adjustment to multiple confounders on any scale.(138)

VI.d. Differential surveillance and follow-up time

In the previous sections we discussed how patients enter a new-user study cohort and how we ensure that patient pre-exposure characteristics are balanced between groups in the absence of baseline randomization. We now need to consider how patients exit a study cohort, an issue that is similarly encountered in RCTs as in RWE studies.(139) While highly-controlled efficacy RCTs spend much effort to keep patients enrolled until the planned study end, in RWE studies we rely on the healthcare system to track patients which is often less than optimal. It is well known that if duration of follow-up differs between treatment groups and if treatment effects vary with duration of use this can lead to meaningful bias.(140)

Three key issues need to be considered to ensure balanced follow-up time. First, in RWE studies with secondary data we need to ensure that the events of interest are captured with the same likelihood in both treatment groups. Patients with better access to care and who see their physician more often are more likely to have diagnoses recorded and medications filled. To improve the odds of balanced outcome surveillance it is common practice to include measures of health service intensity during the pre-exposure phase in the adjustment.(141) Making the reasonable assumption that pre-exposure utilization is predictive of post-exposure utilization this approach will help.

Second, in an as-treated analysis follow-up time is censored at the time the treatment stops plus some exposure effect window. Adherence to treatment is surprisingly bad in clinical practice and one of the contributing factors is affordability of medication, including the level of patient co-payment.(142,143) This is likely the reason why we often see shorter adherence to newer, more costly medications compared with established, often generically available drugs.(4) As above, if treatment effects do not vary with duration of use this is less of an issue. Another pragmatic approach is to limit the follow-up time in both arms to a duration that is well observed among the new drug users with the shorter duration of use.

Third, competing risks, usually death from an unrelated cause, is an issue encountered in all therapeutic studies with time-to-event data, including RCTs.(144,145) Patients treated with an anti-diabetic medication may die of CV complications before experiencing hypoglycemia. A pragmatic solution is the bundle endpoints into a composite that includes death. Statistical modelling techniques are available for time-to-event data that consider competing risks.(146,147)
VI.e. Subgroup analyses and treatment effect modification

When baseline risk of an outcome varies within a population, e.g. the risk of CV events is higher in patients with longer duration of diabetes, the effect of a treatment on that outcome will vary; it will vary either on the additive scale (risk differences) or on the multiplicative scale (risk/hazard ratios). Identifying patient subgroups with increased effectiveness helps personalize treatment plans, e.g. SGLT-2i’s and their CV benefits are more pronounced among patients with pre-existing CV conditions. Conversely, identifying patient segments with increased risk of adverse events will be useful to effectively manage such harms, e.g. the relative risk of lower-limb amputations for canagliflozin compared to other second-line treatments may be highest among older patients with existing CV disease.(148)

Large real-world data sources provide the opportunity to stratify an analysis by many factors that are relevant to prescribers and their patients. Measurement issues are sometimes in the way of identifying mutually exclusive subgroups. General recommendations for studying heterogeneous treatment effects apply to RWE as to RCTs.(149) A particular concern remains post-hoc screening for effect modification that may produce false positive findings despite fairly conservative statistical test for interaction. It can be remedied with Bayesian shrinkage estimation that imposes a less or more strong prior probability of no effect modification.(150) Signals of effect modification should be confirmed in subsequent studies using other data sources and require close collaboration between clinical science and statistical modeling.(151)

VI.f Sensitivity analyses

A series of sensitivity analyses can help investigators better determine how robust a study’s findings are. In Table 3 we list several frequently used sensitivity analyses in pharmacoepidemiology. Some of those are variations of key study design parameters that can be incorporated in the planning and implementation of the study, and others are “back-end” analyses that are based on the actual study finding.(16)

An important yet underutilized tool for detecting the impact of unobserved confounding on the validity of findings in non-interventional studies is quantitative bias analysis. Basic bias analyses of residual confounding seek to determine how strong and how imbalanced an unobserved confounder would have to be among exposure groups to explain the observed effect.(17,105) Lash and Fox proposed a comprehensive approach that considers several systematic errors simultaneously, allowing sensitivity analyses for confounding, misclassification, and selection bias in one process.(90)

In most pharmacoepidemiology studies based on large patient databases, new users of the drugs of interest are identified empirically by a drug dispensing that was not preceded by an earlier use of the same drug. A typical choice of the washout window duration in U.S. claims data is six months; a sensitivity analyses could extend the window to nine or twelve months. Although increasing the length of the washout window increases the likelihood that patients are truly new users, it may also reduce the number of patients eligible for the study in left-censored data, because the longer eligibility requirement excludes some patients.(152) This trade-off is particularly worth noting in health plans or provider networks with high enrollee turnover. Analogously, the covariate assessment window could be extended in sensitivity analyses in order to capture more information that defines the study patients’ health state.

There is often uncertainty about the correct definition of the exposure risk window based on the clinical pharmacology of the study agent and the current understanding of the biology and physiology. When studying the effect of pioglitazone there was a debate whether pioglitazone was a
tumor inducer, which would require a very long exposure risk window until the cancer may become clinically apparent versus a tumor promoter which could have a shorter risk window.(153) Varying the exposure risk window is therefore insightful and easy to accomplish in cohort studies.(116)

Another sensitivity analysis concerns the potential for informative censoring. Patients change and discontinue treatment because there was a disappointing effect or they experience early signs of a side effect. The more strongly such non-persistence is associated with the outcome, the more strongly biased an as-treated analysis, which censors at the point of discontinuation. Alternatively, the as-started analysis follows all patients for a fixed time period, disregarding any changes in treatment status over time. It will avoid informative censoring but suffers bias as a consequence of increasing exposure mischaracterization with time. In most but not all cases, such mischaracterization will bias effects towards the null, similar to intention-to-treat analyses in randomized trials. Viewed separately, these two analysis types trade different biases, and together, they bound a range of plausible effect estimates.

VII. Study transparency

For RWE to have maximum impact on how we treat patients with diabetes, it must not only be valid but also accepted as valid by decision makers. However, a blanket acceptance of all RWE that reaches decision makers is unlikely. As with randomized controlled trials, we need to provide decision makers with unambiguous reporting of RWE study conduct, tools to facilitate efficient review, and guidance on how to assess the validity of results. Decision-makers need to be able to fully understand—and in some cases, reproduce and robustness-check—RWE studies, to build the necessary confidence in using such evidence to inform high-stakes decisions.(44)

VII.a. Study registration

It is a clear advantage that RWE studies are conducted with previously collected information that the healthcare system has produced during the provision of care. The fact that all information is already collected, including information on the study endpoints raises concerns regarding the unbiased conduct of RWE research. Financial and other incentives may tempt investigators to cherry-pick results or adjust their analysis plan if results are not what they had expected them to be.(154) In prospective RCTs this is less of an issue as detailed protocols and statistical analysis plans are deposited before the first patient is recruited. Registration of hypothesis-evaluating treatment-effectiveness studies, providing the specifications for a priori planned analyses along with an audit trail of revisions to the plan, has been proposed as an important step toward improving transparency and confidence in RWE studies of effectiveness.(26) Several options are available for registration of non-interventional studies, including the EU Post-authorizations study register, hosted by the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP.eu), and ClinicalTrials.gov hosted by the National Library of Medicine.(155)

VII.b. Transparent reporting and reproducibility

The first step to understanding the validity of RWE studies is to learn what exactly was done in a given study. How were the data curated, and what transformations were performed upon the longitudinal streams of healthcare encounters contained in the source data to identify the study population, to define drug exposure, to ascertain outcomes, and to balance treatment groups in the absence of randomization? We appreciate RCTs not only because of the power of baseline randomization but also because they can provide clear, simple answers to all of the above. Decision makers see the complexity of RWE and the lack of transparency in reporting as a major barrier to using RWE study findings for decision making.(23)
A joint task force between the *International Society of Pharmacoepidemiology* (ISPE) and the *International Society of Pharmacoeconomics and Outcomes Research* (ISPOR), two professional societies that focus on RWE analyses, defined a catalogue of specific design and implementation parameters. Unambiguous reporting of these parameters was deemed by the large, international group of stakeholders as important to enable reproducible study findings and facilitate validity assessment. Similar items are also reflected in the RECORD-PE reporting guideline.

Once study parameters are accurately recorded, studies become reproducible, enabling the evaluation of their validity before decisions are made. Wang et al. demonstrated the reproducibility of 32 database studies and, in a follow-on study, reasonable reproducibility of 150 published database studies if study parameters are fully and accurately recorded. What is customary in the world of RCTs still needs improvement in RWE.

### VII.c. Software tools

The field of RWE has made tremendous progress. RWE analyses, particularly the arranging of the data stream and the longitudinal measurements, can be complex and the statistical programming code in any line programming language can become unreadable for anybody but the original programmer. This makes it difficult and often impossible to check whether the study plan that may have been described in great detail in a protocol was implemented as intended. Investigators have resorted to double programming, that is, two programmers work off the identical protocol. In most cases we find different results, not often meaningfully different but always numerically different results. A key reason for these discrepancies is that any protocol leaves minor ambiguities and well-intentioned and well-trained programmers may differ in how they interpret a specific instruction and make decisions on the go in order for the program to run. Some of these decisions can substantially impact findings.

Software platforms that are validated procedures to intake raw healthcare data, line them up longitudinally, allow transparent implementations of measurements, study designs, and statistical analyses have proven to reduce such ambiguity and providing the highest transparency possible in the study conduct. Software tools can support causal study designs by incorporating target trial thinking and provide guard rails to investigators so to avoid known design biases. Such commercially and publicly available software platforms are increasingly used and will contribute to high-quality RWE.

### VIII. Putting it all together: A didactic flow diagram

There is an abundance of high-level guidance documents from professional societies and regulators that try to help researchers in the planning and conduct of RWE studies. The development of robust RWE studies based on the target trial paradigm is a logical and linear process (see Figure 5) that should provide needed clarity in planning and evaluating studies. Tools were developed to help investigators design studies in transparent and logical ways, and there are comprehensive software platforms available that provide guardrails in the RWE study design process and provide complete transparency, including audit trails and reporting functions.

In Figure 11 we provide a didactic flow diagram that presents a range of considerations in conducting a typical comparative effectiveness study. Clearly not all aspects of a study implementation can be condensed into such a diagram; however, it provides a sense of the topic areas that need to be addressed in a sequential way. In this example we illustrate a propensity score analysis to achieve balance in pre-exposure patient characteristics between treatment groups,
because that technique is most frequently used. It can of course be replaced by a range of other methods.(163) Specifically, the measurement layer described in Table 5 is abbreviated in the flow diagram. This is because it is highly dependent on the medical context and the data source in question, as explained in earlier sections. We therefore simplify and assume acceptable measurement characteristics for all relevant variables.

The next, more-detailed, level of help in conducting studies would be to consult the recommendations of the ISPE/ISPOR task force on the parameters that need to be defined to implement a study in a reproducible way.(25) It provides a detailed list of study parameters that need to be considered with synonyms, explanations, and examples, all in a logical order similar to the three layers in Figure 5 or the flow diagram in Figure 11. Graphical illustrations of longitudinal study designs complement it.(47)

While didactically helpful, none of this can replace a formal educational curriculum in causal inference, epidemiologic study designs, and biostatistical methods. And blessed are those who have a wise mentor who keeps them from straying off the narrow path to causal scientific insights and provides them with the support needed to move the field forward.

IX. Concluding remarks

RWE has developed rapidly over the past two decades and has made many breakthroughs in recent years, particularly in understanding treatment options in diabetes care. It is here to stay as an integral part of a comprehensive evidence-development framework complementing RCTs, basic science, and many other elements of scientific discovery.

For a chronic disease like diabetes with many treatment options that are combined and staged in complex ways during disease progression, there will never be enough randomized trials to answer all relevant treatment-related questions and RWE will increasingly complement RCTs to inform physicians, regulators, guideline writers, payers, and ultimately patients in their therapeutic decision-making process. The ever larger, more detailed, and complete digital data resources that become available to researchers will push RWE into areas that we currently avoid because of imprecise measurement of outcome or confounding factors.

Any implementation of RWE studies needs to be driven by the desire to come to causal conclusions of the treatments under study. The target trial thinking provides clarity in the planning and interpretation of RWE studies and paired with the many new tools that epidemiology and biostatistics offer will guide investigators towards that goal. However we should conclude with words of caution. “[The] complexity of actual context [in medicine] prohibits anything approaching complete causal modeling; the models actually used are never entirely coherent with that context, and formal analyses can only serve as thought experiments within informal guidelines. A cautious scientists will thus reserve judgement and treat no methodology as correct or absolute, but will instead examine data from multiple perspectives, taking statistical models for what they are: semi-automated algorithms which are error-prone without sophisticated human input from both methodologists and content experts.”(164)
Figures and Tables

Figure 1: Data sources used by RWE studies

EHR = Electronic health records; NDI = National Death Index; PRO = Patient-reported outcomes;
Adapted from: Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real world data analyses for regulatory decision making. Clin Pharm Ther 2019;105:867-77.

Figure 2: Contributions of RWE for regulatory and coverage decision making

Adapted from: Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real world data analyses for regulatory decision making. Clin Pharm Ther 2019;105:867-77.

Figure 3: From longitudinal electronic healthcare data to a causal cohort study design

Adapted from: Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. Clin Epidemiol, 2018;10;771-88.

Figure 4: Illustration of typical longitudinal study design choices in pharmacoepidemiology*

* The diagrams use a comparative analysis of new users of pioglitazone versus new users of rosiglitazone on some health outcomes to illustrate how time windows are used to identify key markers and variables.

Figure 5: From real-world data to real-world evidence

Figure 6: The study question and sources of treatment exposure variation guide design choices

Adapted from: Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Safety 2010;19:858-68.

Figure 7: Schematic of a parallel group RCT and the corresponding cohort design

Figure 8: The prediction of the CAROLINA RCT by a RWE study

Figure 9: Immortal time bias in diabetes RWE studies illustrated in three hypothetical patients*

T2DM = type-2 diabetes mellites; SGLT-2i = Sodium-glucose transporter-2 inhibitors; oGLD = other glucose-lowering drugs

Figure 10: Checking balance of unmeasured pre-exposure covariates in a subsample of patients with enriched clinical data

EHR = electronic health records

Figure 11: A flow diagram of considerations for a typical RWE study using healthcare databases*

*Modified after Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Safety 2010;19:858-68.
Table 1: Measurement characteristics that inform RWE study validity and often-quoted proxies

| Study features: | Examples of ways to improve measurement characteristics | Typical proxies for data quality in secondary data | Actual measurement characteristics* |
|----------------|----------------------------------------------------------|-------------------------------------------------|-----------------------------------|
| 1) Study population, subgroups | Require two diagnosis codes to increase specificity of underlying condition | Prior experience with a data source, publications | Binary data, e.g. diagnostic codes present: |
| 2) Exposure measurement | Use dispensing information instead of prescribing data to increase completeness | Availability of validation studies | • Sensitivity |
| 3) Outcome measurement | Use serious events, e.g. that require hospitalizations to increase specificity of outcome measurement | Detailed documentation of data generation mechanism | • Specificity |
| 4) Confounder measurement | Screen a wide range of potential confounders and their proxies to limit unobserved confounding | Detailed description of data curating process | • PPV |
| | | Detailed description of mapping to medical constructs (if any) | Continuous data, e.g. lab test values: |
| | | Documentation of coding shift over time | • % missing |
| | | | Mean squared deviation |
| | | | Time-to-event: |
| | | | • Accuracy of onset |

Adapted from: Franklin JM, Glynn RI, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real world data analyses for regulatory decision making. Clin Pharm Ther 2019;105:867-77.
Table 2: Balance in relevant pre-exposure patient characteristics before and after PS matching in a typical new-user, active-comparator cohort study

| Baseline characteristics | Linagliptin (N = 24,842) | Glimepiride (N = 139,334) | St. diff. | Linagliptin (N = 24,131) | Glimepiride (N = 24,131) | St. diff. |
|--------------------------|--------------------------|---------------------------|-----------|--------------------------|--------------------------|-----------|
| Age; mean (s.d.)         | 70.32 (7.76)             | 71.18 (7.72)              | -0.11     | 70.40 (7.76)             | 70.42 (7.71)             | 0.00      |
| Male; n (%)              | (47.8%)                  | (68,926 (49.5%)           | -0.03     | (47.7%)                  | (11,519)                 | 0.00      |
| White race; n (%)        | (72.0%)                  | (71,226 (79.8%)           | -0.21     | (71.8%)                  | (11,485)                 | -0.01     |
| **Burden of comorbidities** |                          |                           |           |                          |                           |           |
| Combined comorbidity score; mean (s.d.) | 1.49 (1.27)       | 1.21 (1.21)               | 0.23      | 1.48 (1.26)              | 1.47 (1.26)              | 0.01      |
| **Diabetes-related complications** |                          |                           |           |                          |                           |           |
| Diabetic nephropathy; n (%) | 1,771 (7.1%)          | 6,350 (4.6%)              | 0.11      | 1,655 (6.9%)             | 1,646 (6.8%)             | 0.00      |
| Diabetic retinopathy; n (%) | 1,172 (4.7%)         | 5,754 (4.1%)              | 0.03      | 1,132 (4.7%)             | 1,103 (4.6%)             | 0.00      |
| Diabetes with ophthalmic conditions; n (%) | 656 (2.6%)          | 3,363 (2.4%)              | 0.01      | 632 (2.6%)               | 625 (2.6%)               | 0.00      |
| Diabetic neuropathy; n (%) | 2,765 (11.1%)         | 14,093 (10.1%)            | 0.03      | 2,663 (11.0%)            | 2,656 (11.0%)            | 0.00      |
| Diabetes with peripheral circ. disorders; n (%) | 1,099 (4.4%)         | 5,309 (3.8%)              | 0.03      | 1,050 (4.4%)             | 1,039 (4.3%)             | 0.00      |
| Hypoglycemia; n (%)      | 708 (2.9%)              | 3,212 (2.3%)              | 0.04      | 684 (2.8%)               | 695 (2.9%)               | -0.01     |
| Hyperglycemia; n (%)     | 997 (4.0%)              | 3,856 (2.8%)              | 0.07      | 961 (4.0%)               | 968 (4.0%)               | 0.00      |
| Disorders of fluid electrolyte balance; n (%) | 1,114 (4.5%)         | 5,333 (3.8%)              | 0.04      | 1,083 (4.5%)             | 1,097 (4.5%)             | 0.00      |
| **Pre-exposure diabetes therapy** |                          |                           |           |                          |                           |           |
| Any use of metformin; n (%) | 16,477 (66.3%)        | 88,730 (63.7%)            | 0.05      | 16,009 (66.9%)           | 16,136 (66.9%)           | -0.01     |
| Any use of sulfonylureas; n (%) | 8,043 (32.4%)         | 21,751 (15.6%)            | 0.40      | 7,458 (30.9%)            | 7,341 (30.4%)            | 0.01      |
| Any use of meglitinides; n (%) | 358 (1.4%)           | 774 (0.6%)                | 0.08      | 341 (1.4%)               | 293 (1.2%)               | 0.02      |
| Any use of alpha-glucosidase inhibitors; n (%) | 84 (0.3%)            | 240 (0.2%)                | 0.02      | 74 (0.3%)                | 92 (0.4%)                | -0.02     |
| **Comorbidities at baseline** |                          |                           |           |                          |                           |           |
| Ischemic heart disease; n (%) | 6,194 (24.9%)         | 33,882 (24.3%)            | 0.01      | 6,016 (24.9%)            | 6,042 (25.0%)            | 0.00      |
| Heart failure; n (%)     | 915 (3.7%)              | 5,144 (3.7%)              | 0.00      | 896 (3.7%)               | 865 (3.6%)               | 0.01      |
| Hypertension; n (%)      | 21,592 (86.9%)          | 114,186 (82.0%)           | 0.14      | 20,940 (86.8%)           | 20,985 (87.0%)           | -0.01     |
| Hyperlipidemia; n (%)    | 19,154 (77.1%)          | 100,099 (71.8%)           | 0.12      | 18,581 (77.0%)           | 18,583 (77.0%)           | 0.00      |
| **Measures of healthcare utilization** |                          |                           |           |                          |                           |           |
| Hospitalization within prior 30 days; n (%) | 260 (1.0%)           | 1,774 (1.3%)              | -0.03     | 255 (1.1%)               | 284 (1.2%)               | -0.01     |
| No. office visits; mean (s.d.) | 4.83 (3.62)        | 4.09 (3.34)               | 0.21      | 4.77 (3.56)              | 4.81 (3.78)              | -0.01     |

PS = propensity score; St. diff. = standardized difference; s.d. = standard deviation

Data from Patorno E, Schneeweiss S, Gopalakrishnan C, Martin D, Franklin JM. Using Real-World Data to Predict Findings of an Ongoing Phase IV Cardiovascular Outcome Trial: Cardiovascular Safety of Linagliptin Versus Glimepiride. Diabetes Care. 2019;42:2204-10.
| **“Front end”: Design choices** | **Some sensitivity analyses:** |
|---------------------------------|-------------------------------|
| 1) Incomplete covariate assessment | Extend covariate assessment window further to the past |
| 2) Reverse causation | Include induction (lag) time |
| 3) Mis-specified follow-up model (handling right censoring) | “As-treated” vs. “As-started” effects |
| 4) Mis-specified exposure-risk-window (ERW) | Vary length of the exposure effect window |
| 5) Differential surveillance bias | Adjust for healthcare utilization intensity |

| **“Back end”: Quantitative bias analysis** | **Some sensitivity analyses:** |
|-------------------------------------------|-------------------------------|
| 6) Imperfect outcome measurement | Quantitative bias analysis |
| 7) Imperfect exposure measurement (see #4) | Quantitative bias analysis |
| 8) Residual confounding due to imperfect measurement | Quantitative bias analysis |
References

1. Rosenstock J, Kahn SE, Johansen OE, Zinman B, Espeland MA, Woerle HJ, Pfarr E, Keller A, Mattheus M, Baanstra D, Meinicke T, George JT, von Eynatten M, McGuire DK, Marx N, Investigators C. Effect of Linagliptin vs Glimepiride on Major Adverse Cardiovascular Outcomes in Patients With Type 2 Diabetes: The CAROLINA Randomized Clinical Trial. JAMA. 2019.

2. Neal B, Perkovic V, Mahaffey KW, de Zeeuw D, Fulcher G, Erondou N, Shaw W, Law G, Desai M, Matthews DR, Group CPC. Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes. N Engl J Med. 2017;377(7):644-657.

3. Wiviott SD, Raz I, Bonaca MP, Mosenzon O, Kato ET, Cahn A, Silverman MG, Zelniker TA, Jonasson C, Wintzell V, Melbye M, Svanström H. Use of sodium glucose cotransporter 2 inhibitors and risk of major cardiovascular events and heart failure: Scandinavian register based cohort study. BMJ. 2019;366:l4772.

4. Patorno E, Goldfine AB, Schneeweiss S, Everett BM, Glynn RJ, Liu J, Kim SC. Cardiovascular outcomes associated with canagliflozin versus other non-gliflozin antidiabetic drugs: population based cohort study. BMJ. 2018;360:k119.

5. Pasternak B, Ueda P, Eliasson B, Svensson A-M, Franden S, Gudbjornsdottir S, Hveem K, Jonasson C, Wintzell V, Melbye M, Svansström H. Use of sodium glucose cotransporter 2 inhibitors and risk of major cardiovascular events and heart failure: Scandinavian register based cohort study. BMJ. 2019;366:l4772.

6. Patorno E, Pawar A, Franklin JM, Najafzadeh M, Denz-Luyet A, Brodovicz KG, Sambevski S, Bessette LG, Santiago Ortiz AJ, Kulendorff M, Schneeweiss S. Empagliflozin and the Risk of Heart Failure Hospitalization in Routine Clinical Care. Circulation. 2019;139(25):2822-2830.

7. Filion KB, Lix LM, Yu OH, Dell’Aniello S, Drousos A, Shah BR, St-Jean A, Fisher A, Tremblay E, Bugden SC, Alessi-Severini S, Ronksley PE, Hu N, Dormuth CR, Ernst P, Suisa S, Canadian Network for Observational Drug Effect Studies. Sodium glucose cotransporter 2 inhibitors and risk of major adverse cardiovascular events: multi-database retrospective cohort study. BMJ. 2020;370:m3342.

8. Grodstein F, Stampfer MJ, Manson JE, Colditz GA, Willett WC, Rosner B, Speizer FE, Hennekens CH. Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. N Engl J Med. 1996;335(7):453-461.

9. Hernandez MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, Manson JE, Robins JM. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008;19(6):766-779.

10. Stamper MJ, Hennekens CH, Manson JE, Colditz GA, Rosner B, Willett WC. Vitamin E consumption and the risk of coronary heart disease. N Engl J Med. 1993;328(20):1444-1449.

11. Rimm EB, Stampfer MJ, Ascherio A, Giovannucci E, Colditz GA, Willett WC. Vitamin E consumption and the risk of coronary heart disease in men. N Engl J Med. 1993;328(20):1450-1456.

12. Heart Outcomes Prevention Evaluation Study I, Yusuf S, Dagenais G, Pogue J, Bosch J, Sleight P. Vitamin E supplementation and cardiovascular events in high-risk patients. N Engl J Med. 2000;342(3):154-160.

13. Chan KA, Andrade SE, Boles M, Buist DS, Chase GA, Donahue JG, Goodman MJ, Gurwitz JH, LaCroix AZ, Platt R. Inhibitors of hydroxymethylglutaryl-coenzyme A reductase and risk of fracture among older women. Lancet. 2000;355(9222):2185-2188.

14. Heart Protection Study Collaborative G. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet. 2002;360(9326):7-22.
15. Framework for FDA's Real World Evidence Program. 2018. [https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RealWorldEvidence/UCM627769.pdf]. Accessed 1/31/2019

16. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol.* 2005;58(4):323-337.

17. Patorno E, Gopalakrishnan C, Franklin JM, Brodovitz KG, Masso-Gonzalez E, Bartels DB, Liu J, Schneeweiss S. Claims-based studies of oral glucose-lowering medications can achieve balance in critical clinical variables only observed in electronic health records. *Diabetes Obes Metab.* 2018;20(4):974-984.

18. Bykov K, He M, Franklin JM, Garry EM, Seeger JD, Patorno E. Glucose-lowering medications and the risk of cancer: a methodological review of studies based on real-world data. *Diabetes Obes Metab.* 2019;21(9):2029-2038.

19. Suissa S. Metformin to Treat Cancer: Misstep in Translational Research from Observational Studies. *Epidemiology.* 2017;28(3):455-458.

20. Suissa S. Lower Risk of Death With SGLT2 Inhibitors in Observational Studies: Real or Bias? *Diabetes Care.* 2018;41(1):6-10.

21. Suissa S. Reduced Mortality With Sodium-Glucose Cotransporter-2 Inhibitors in Observational Studies: Avoiding Immortal Time Bias. *Circulation.* 2018;137(14):1432-1434.

22. Patorno E, Schneeweiss S, Gopalakrishnan C, Martin D, Franklin JM. Using Real-World Data to Predict Findings of an Ongoing Phase IV Cardiovascular Outcome Trial: Cardiovascular Safety of Linagliptin Versus Glimepiride. *Diabetes Care.* 2019;42(12):2204-2210.

23. Malone DC, Brown M, Hurwitz JT, Peters L, Graff JS. Real-World Evidence: Useful in the Real World of US Payer Decision Making? How? When? And What Studies? *Value Health.* 2018;21(3):326-333.

24. Husereau D, Nason E, Ahuja T, Nikai E, Tsakonas E, Jacobs P. Use of Real-World Data Sources for Canadian Drug Pricing and Reimbursement Decisions: Stakeholder Views and Lessons for Other Countries. *Int J Technol Assess Health Care.* 2019;35(3):181-188.

25. Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, Gagne JJ, Gini R, Klungel O, Mullins CD, Nguyen MD, Rassen JA, Smeeth L, Sturkenboom M, joint ISTFoRWEiHCDM. Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0. *Pharmacoepidemiology and Drug Saf.* 2017;26(9):1018-1032.

26. Berger ML, Sox H, Willke RJ, Brixner DL, Eichler HG, Goetttsch W, Madigan D, Makady A, Schneeweiss S, Tarricone R. 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. *Pharmacoepidemiology and Drug Safety.* 2017;26(9):1033-1039.

27. Orsini LS, Berger M, Crown W, Daniel G, Eichler HG, Goetttsch W, Graff J, Guerino J, Jonsson P, Lederer NM, Monz B, Mullins CD, Schneeweiss S, Brunt DV, Wang SV, RJ. W. Improving Transparency to Build Trust in Real-World Secondary Data Studies for Hypothesis Testing-Why, What, and How: Recommendations and a Road Map from the Real-World Evidence Transparency Initiative. *Value Health.* 2020;23:1128-1136.

28. Patorno E, Schneeweiss S, Wang SV. Transparency in real-world evidence (RWE) studies to build confidence for decision-making: Reporting RWE research in diabetes. *Diabetes Obes Metab.* 2020;22 Suppl 3:34-59.

29. Wang SV, Verpillat P, Rassen JA, Patrick A, Garry EM, Bartels DB. Transparency and Reproducibility of Observational Cohort Studies Using Large Healthcare Databases. *Clin Pharmacol Ther.* 2016;99(3):325-332.

30. Franklin JM, Pawar A, Martin D, Glynn RJ, Levenson M, Temple R, Schneeweiss S. Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project. *Clin Pharmacol Ther.* 2020;107(4):817-826.
31. Ray WA, Stein CM, Daugherty JR, Hall K, Arbogast PG, Griffin MR. COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. *Lancet*. 2002;360(9339):1071-1073.

32. Southworth MR, Reichman ME, Unger EF. Dabigatran and postmarketing reports of bleeding. *The New England journal of medicine*. 2013;368(14):1272-1274.

33. Graham DI, Reichman ME, Wernecke M, Zhang R, Southworth MR, Levenson M, Sheu TC, Mott K, Goulding MR, Houstoun M, MaCurdy TE, Worrall C, Kelman JA. Cardiovascular, bleeding, and mortality risks in elderly Medicare patients treated with dabigatran or warfarin for nonvalvular atrial fibrillation. *Circulation*. 2015;131(2):157-164.

34. Platt RW, Platt R, Brown JS, Henry DA, Klungel OH, Suissa S. How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias. *Pharmacoepidemiol Drug Saf*. 2019.

35. Cefalu WT, Kaul S, Gerstein HC, Holman RR, Zinman B, Skyler JS, Green JB, Buse JB, Inzucchi SE, Leiter LA, Raz I, Rosenstock I, Riddle MC. Cardiovascular, bleeding, and mortality risks in elderly Med¡care patients treated with dabigatran or warfarin for nonvalvular atrial fibrillation. *Circulation*. 2015;131(2):157-164.

36. Platt RW, Platt R, Brown JS, Henry DA, Klungel OH, Suissa S. How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias. *Pharmacoepidemiol Drug Saf*. 2019.

37. Fralick M, Schneeweiss S, Patorno E. Risk of Diabetic Ketoacidosis after Initiation of an SGLT2 Inhibitor. *N Engl J Med*. 2017;376(23):2300-2302.

38. US_Food_and_Drug_Administration. CFR - Code of Federal Regulations Title 21. Vol 20202019.

39. Fralick M, Kesselheim AS, Avorn J, Schneeweiss S. Use of Health Care Databases to Support Supplemental Indications of Approved Medications. *JAMA internal medicine*. 2018;178(1):55-63.

40. Garry EM, Schneeweiss S, Eapen S, Petruski-Ivleva N, Cheever E, Murk W, Rajan A, Rassen JA, Gambino D, Jang H, Rubin E, Jan S. Actionable Real-World Evidence to Improve Health Outcomes and Reduce Medical Spending Among Risk-Stratified Patients with Diabetes. *J Manag Care Spec Pharm*. 2019;25(12):1442-1452.

41. Franko EM, Schneeweiss S, Eapen S, Petruski-Ivleva N, Cheever E, Murk W, Rajan A, Rassen JA, Gambino D, Jang H, Rubin E, Jan S. Actionable Real-World Evidence to Improve Health Outcomes and Reduce Medical Spending Among Risk-Stratified Patients with Diabetes. *J Manag Care Spec Pharm*. 2019;25(12):1442-1452.

42. Wang SV, Schneeweiss S. START-RWE: A structured template for planning and reporting on the implementation real-world evidence studies *Br Med J*. 2020;In press.

43. Shadish WR, Clark MH, Steiner PM. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *J Am Stat Assoc*. 2008;103(484):1334-1343.

44. Schneeweiss S. Real-World Evidence of Treatment Effects: The Useful and the Misleading. *Clin Pharmacol Ther*. 2019;106(1):43-44.

45. Forbes SP, Dahabreh IJ. Benchmarking Observational Analyses Against Randomized Trials: a Review of Studies Assessing Propensity Score Methods. *J Gen Intern Med*. 2020;35(5):1396-1404.

46. Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, Pawar A, Bessette LG, Lee H, Garry EM, Gautam N, Schneeweiss S. Emulating Randomized Clinical Trials with Nonrandomized Real-World Evidence Studies: First Results from the RCT DUPLICATE Initiative. *Circulation*. 2020.

47. Schneeweiss S, Rassen JA, Brown JS, Rothman KJ, Happe L, Arlett P, Dal Pan G, Goettsch W, Murk W, Wang SV. Graphical Depiction of Longitudinal Study Designs in Health Care Databases. *Ann Intern Med*. 2019.

48. VanderWeele TJ. Principles of confounder selection. *European journal of epidemiology*. 2019;34(3):211-219.
49. Gokhale M, Sturmer T, Buse JB. Real-world evidence: the devil is in the detail. Diabetologia. 2020;63(9):1694-1705.

50. Wettermark B, Zoega H, Furu K, Korhonen M, Halls J, Norgaard M, Almardsdottir A, Andersen M, Andersson Sundell K, Bergman U, Helin-Salmivaara A, Hoffmann M, Kieler H, Martikainen J, Mortensen M, Petzold M, Wallach-Kildemoes H, Wallin C, Sorensen H. The Nordic prescription databases as a resource for pharmacoepidemiological research—a literature review. Pharmacoepidemiol Drug Saf. 2013;22(7):691-699.

51. ISO/TR. Health Informatics - Electronic Health Record Definition, Scope, and Context. Vol 20514 2005.

52. Mandl KD, Kohane IS. Escaping the EHR trap—the future of health IT. N Engl J Med. 2012;366(24):2240-2242.

53. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc. 2006;13(2):121-126.

54. The Editors Of The Lancet G. Learning from a retraction. Lancet. 2020;396(10257):1056.

55. Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol. 2000;29(4):722-729.

56. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol. 2003;158(9):915-920.

57. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010;19(8):858-868.

58. Johnson ES, Bartman BA, Briesacher BA, Fleming NS, Gerhard T, Kornegay CJ, Nourjah P, Sauer B, Schumock GT, Sedrakyan A, Sturmer T, West SL, Schneeweiss S. The incident user design in comparative effectiveness research. Pharmacoepidemiology and Drug Safety. 2013;22(1):1-6.

59. Faillie JL, Yu OH, Yin H, Hillaire-Buys D, Barkun A, Azoulay L. Association of Bile Duct and Gallbladder Diseases With the Use of Incretin-Based Drugs in Patients With Type 2 Diabetes Mellitus. JAMA internal medicine. 2016;176(10):1474-1481.

60. Maclure M. 'Why me?' versus 'why now?'—differences between operational hypotheses in case-control versus case-crossover studies. Pharmacoepidemiology and Drug Saf. 2007;16(8):850-853.

61. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol. 1991;133(2):144-153.

62. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology. 2006;17(3):268-275.

63. Winkelmann WC, Setoguchi S, Levin R, Solomon DH. Comparison of cardiovascular outcomes in elderly patients with diabetes who initiated rosiglitazone vs pioglitazone therapy. Arch Intern Med. 2008;168(21):2368-2375.

64. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. American journal of epidemiology. 2016;183(8):758-764.

65. Hernan MA, Sauer BC, Hernandez-Diaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol. 2016;79:70-75.

66. Suissa S. Immortal time bias in observational studies of drug effects. Pharmacoepidemiology and Drug Safety. 2007;16(3):241-249.

67. Moran LV, Ongur D, Hsu J, Castro VM, Perlis RH, Schneeweiss S. Psychosis with Methylphenidate or Amphetamine in Patients with ADHD. N Engl J Med. 2019;380(12):1128-1138.
68. Smeth L, Douglas I, Hall AJ, Hubbard R, Evans S. Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials. Br J Clin Pharmacol. 2009;67(1):99-109.
69. Moride Y, Abenhaim L. Evidence of the depletion of susceptibles effect in non-experimental pharmacoepidemiologic research. J Clin Epidemiol. 1994;47(7):731-737.
70. Goodman SN, Schneeweiss S, Baiocchi M. Using Design Thinking to Differentiate Useful From Misleading Evidence in Observational Research. JAMA. 2017;317(7):705-707.
71. Suissa S, Moodie EE, Dell’Aniello S. Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores. Pharmacoepidemiol Drug Saf. 2017;26(4):459-468.
72. Setoguchi S, Warner Stevenson L, Stewart GC, Bhatt DL, Epstein AE, Desai M, Williams LA, Chen CY. Influence of healthy candidate bias in assessing clinical effectiveness for implantable cardioverter-defibrillators: cohort study of older patients with heart failure. BMJ. 2014;348:g2866.
73. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol. 2008;167(4):492-499.
74. Kosiborod M, Cavender MA, Fu AZ, Wilding JP, Khunti K, Holl RW, Norhammar A, Birkeland KI, Jorgensen ME, Thuresson M, Aya N, Bodegard J, Hammar N, Fenici P, Investigators C-R, Study G. Lower Risk of Heart Failure and Death in Patients Initiated on Sodium-Glucose Cotransporter-2 Inhibitors Versus Other Glucose-Lowering Drugs: The CVD-REAL Study (Comparative Effectiveness of Cardiovascular Outcomes in New Users of Sodium-Glucose Cotransporter-2 Inhibitors). Circulation. 2017;136(3):249-259.
75. Udell JA, Yuan Z, Rush T, Sicignano NM, Galitz M, Rosenthal N. Cardiovascular Outcomes and Risks After Initiation of a Sodium Glucose Cotransporter 2 Inhibitor: Results From the EASEL Population-Based Cohort Study (Evidence for Cardiovascular Outcomes With Sodium Glucose Cotransporter 2 Inhibitors in the Real World). Circulation. 2018;137(14):1450-1459.
76. Schneeweiss S, Suissa S. Discussion of Schuemie et al: "A plea to stop using the case-control design in retrospective database studies". Stat Med. 2019;38(22):4209-4212.
77. Miettinen OS. Estimation of relative risk from individually matched series. Biometrics. 1970;26(1):75-86.
78. Miettinen O. Estimability and estimation in case-referent studies. Am J Epidemiol. 1976;103(2):226-235.
79. Dickerman BA, Garcia-Albeniz X, Logan RW, Denaxas S, Hernan MA. Emulating a target trial in case-control designs: an application to statins and colorectal cancer. Int J Epidemiol. 2020.
80. Wacholder S. Practical considerations in choosing between the case-cohort and nested case-control designs. Epidemiology. 1991;2(2):155-158.
81. Wu JW, Filion KB, Azoulay L, Doll MK, Suissa S. Effect of Long-Acting Insulin Analogs on the Risk of Cancer: A Systematic Review of Observational Studies. Diabetes Care. 2016;39(3):486-494.
82. Moore N, Duret S, Grolleau A, Lassalle R, Barbet V, Duong M, Thurin N, Droz-Perroteau C, Gulmez SE. Previous Drug Exposure in Patients Hospitalised for Acute Liver Injury: A Case- Population Study in the French National Healthcare Data System. Drug Saf. 2019;42(4):559-572.
83. Yu OH, Filion KB, Azoulay L, Patenaude V, Majdan A, Suissa S. Incretin-based drugs and the risk of congestive heart failure. Diabetes Care. 2015;38(2):277-284.
84. Suissa S, Dell’Aniello S, Martinez C. The multitime case-control design for time-varying exposures. Epidemiology. 2010;21(6):876-883.
85. Suissa S. The Quasi-cohort approach in pharmacoepidemiology: upgrading the nested case-control. Epidemiology. 2015;26(2):242-246.
86. Prentice RL, Self SG. Aspects of the use of relative risk models in the design and analysis of cohort studies and prevention trials. Stat Med. 1988;7(1-2):275-287.
87. Onland-Moret NC, van der AD, van der Schouw YT, Buschers W, Elias SG, van Gils CH, Koerselman J, Roest M, Grobbee DE, Peeters PH. Analysis of case-cohort data: a comparison of different methods. J Clin Epidemiol. 2007;60(4):350-355.
88. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. Am J Epidemiol. 1988;128:1198-1206.
89. Collet JP, Schaubel D, Hanley J, Sharpe C, Boivin JF. Controlling confounding when studying large pharmacoepidemiologic databases: a case study of the two-stage sampling design. Epidemiology. 1998;9:309-315.
90. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014;43(6):1969-1985.
91. West SL, Strom BL, Freundlich B, Normand E, Koch G, Savitz DA. Completeness of prescription recording in outpatient medical records from a health maintenance organization. J Clin Epidemiol. 1994;47(2):165-171.
92. West SL, Savitz DA, Koch G, Strom BL, Guess HA, Hartzema A. Recall accuracy for prescription medications: self-report compared with database information. Am J Epidemiol. 1995;142(10):1103-1112.
93. Rothman KJ, Poole C. A strengthening programme for weak associations. Int J Epidemiol. 1988;17(4):955-959.
94. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol. 2004;57(2):131-141.
95. Funk MJ, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. Current epidemiology reports. 2014;1(4):175-185.
96. Kiyota Y, Schneeweiss S, Glynn RJ, Cannuscio CC, Avorn J, Solomon DH. Accuracy of Medicare claims-based diagnosis of acute myocardial infarction: estimating positive predictive value on the basis of review of hospital records. Am Heart J. 2004;148(1):99-104.
97. Choudhry NK, Brennan T, Toscano M, Spettell C, Glynn RJ, Rubino M, Schneeweiss S, Brookhart AM, Fernandes J, Mathew S, Christiansen B, Antman EM, Avorn J, Shrank WH. Rationale and design of the Post-MI FREEE trial: a randomized evaluation of first-dollar drug coverage for post-myocardial infarction secondary preventive therapies. Am Heart J. 2008;156(1):31-36.
98. Ray WA, Chung CP, Murray KT, Hall K, Stein CM. Atypical antipsychotic drugs and the risk of sudden cardiac death. N Engl J Med. 2009;360(3):225-235.
99. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Medical care. 2005;43(5):480-485.
100. Tirschwell DL, Longstreth WT, Jr. Validating administrative data in stroke research. Stroke. 2002;33(10):2465-2470.
101. Ginde AA, Blanc PG, Lieberman RM, Camargo CA, Jr. Validation of ICD-9-CM coding algorithm for improved identification of hypoglycemia visits. BMC Endocr Disord. 2008;8:4.
102. Schelleman H, Bilker WB, Brensinger CM, Wan F, Hennessy S. Anti-infectives and the risk of severe hypoglycemia in users of glipizide or glyburide. Clin Pharmacol Ther. 2010;88(2):214-222.
103. Bobo WV, Cooper WO, Epstein RA, Jr., Arbogast PG, Mounsey J, Ray WA. Positive predictive value of automated database records for diabetic ketoacidosis (DKA) in children and youth exposed to antipsychotic drugs or control medications: a Tennessee Medicaid Study. BMC Med Res Methodol. 2011;11:157.
104. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the Use of Nonrandomized Real-World Data Analyses for Regulatory Decision Making. Clin Pharmacol Ther. 2019;105(4):867-877.
105. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf.* 2006;15(5):291-303.

106. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *Am J Epidemiol.* 1991;134(8):895-907.

107. Vach W, Blettner M. Logistic regression with incompletely observed categorical covariates -- investigating the sensitivity against violation of the missing at random assumption. *Stat Med.* 1995;14(12):1315-1329.

108. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ.* 2009;338:b2393.

109. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology.* 2009;20(4):512-522.

110. Franklin JM, Eddings W, Schneeweiss S, Rassen JA. Incorporating linked healthcare claims to improve confounding control in a study of in-hospital medication use. *Drug Saf.* 2015;38(6):589-600.

111. Qaseem A, Wilt TJ, Kansagara D, Horwitch C, Barry MJ, Forciea MA, Clinical Guidelines Committee of the American College of P. Hemoglobin A1c Targets for Glycemic Control With Pharmacologic Therapy for Nonpregnant Adults With Type 2 Diabetes Mellitus: A Guidance Statement Update From the American College of Physicians. *Ann Intern Med.* 2018;168(8):569-576.

112. Gamble JM, Chibrikov E, Twells LK, Midodzi WK, Young SW, MacDonald D, Majumdar SR. Association of insulin dosage with mortality or major adverse cardiovascular events: a retrospective cohort study. *Lancet Diabetes Endocrinol.* 2017;5(1):43-52.

113. Farmer RE, Ford D, Mathur R, Chaturvedi N, Kaplan R, Smeeth L, Bhaskaran K. Metformin use and risk of cancer in patients with type 2 diabetes: a cohort study of primary care records using inverse probability weighting of marginal structural models. *Int J Epidemiol.* 2019;48(2):527-537.

114. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.* 2000;11(5):561-570.

115. Benner JS, Glynn RJ, Mogun H, Neumann PJ, Weinstein MC, Avorn J. Long-term persistence in use of statin therapy in elderly patients. *JAMA.* 2002;288(4):455-461.

116. Solomon DH, Lunt M, Schneeweiss S. The risk of infection associated with tumor necrosis factor alpha antagonists: making sense of epidemiologic evidence. *Arthritis Rheum.* 2008;58(4):919-928.

117. Schneeweiss S, Patrick AR, Sturmer T, Brookhart MA, Avorn J, Maclure M, Rothman KJ, Glynn RJ. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Med Care.* 2007;45(10 Suppl 2):S131-142.

118. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol.* 2006;163(12):1149-1156.

119. AM Walker AM, Patrick A, Lauer M, Hornbrook M, Marin M, Platt R, Roger V, Stang P, Schneeweiss S. Tool for Assessing the Feasibility of Comparative Effectiveness Research. *Comp Effect Res* 2013;3:11-20.

120. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. *Stat Med.* 2014;33(10):1685-1699.

121. Desai RJ, Rothman KJ, Bateman BT, Hernandez-Diaz S, HuybrechtsKF. A Propensity-score-based Fine Stratification Approach for Confounding Adjustment When Exposure Is Infrequent. *Epidemiology.* 2017;28(2):249-257.
122. Wooldridge JM. Econometric analysis of cross section and panel data. Cambridge, Mass.: MIT Press.
123. Gelman A, Carlin J, Stern H, Rubin D. Bayesian Data Analysis. New York: Chapman Hall.
124. Suissa S, Blais L, Ernst P. Patterns of increasing beta-agonist use and the risk of fatal or near-fatal asthma. The European respiratory journal. 1994;7(9):1602-1609.
125. Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. Variable Selection for Confounding Adjustment in High-dimensional Covariate Spaces When Analyzing Healthcare Databases. Epidemiology. 2017;28(2):237-248.
126. Karim ME, Pang M, Platt RW. Can We Train Machine Learning Methods to Outperform the High-dimensional Propensity Score Algorithm? Epidemiology. 2018;29(2):191-198.
127. Zhou M, Wang SV, Leonard CE, Gagne JJ, Fuller C, Hamp C, Archieacon P, Toh S, Iyer A, Woodworth TS, Cavagnaro E, Panozzo CA, Axtman S, Carnahan RM, Chrischilles EA, Hennessy S. Sentinel Modular Program for Propensity Score Matched Cohort Analyses: Application to Glyburide, Glipizide, and Serious Hypoglycemia. Epidemiology. 2017;28(6):838-846.
128. Gangji AS, Cukierman T, Gerstein HC, Goldsmith CH, Clase CM. A systematic review and meta-analysis of hypoglycemia and cardiovascular events: a comparison of glyburide with other secretagogues and with insulin. Diabetes Care. 2007;30(2):389-394.
129. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. Clinical epidemiology. 2018;10:771-788.
130. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Instrumental variable analysis of secondary pharmacoepidemiologic data. Epidemiology. 2006;17(4):373-374.
131. Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. N Engl J Med. 2008;358(8):771-783.
132. Gokhale M, Buse JB, DeFilippo Mack C, Jonsson Funk M, Lund J, Simpson RJ, Sturmer T. Calendar time as an instrumental variable in assessing the risk of heart failure with antihyperglycemic drugs. Pharmacoepidemiol Drug Saf. 2018;27(8):857-866.
133. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf. 2010;19(6):537-554.
134. Stukel TA, Fisher ES, Wennberg DE,Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. Jama. 2007;297(3):278-285.
135. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. Int J Biostat. 2007;3(1):14.
136. Brookhart MA, Rassen JA, Schneeweiss S. A Practical Guide to the Use of Instrumental Variable Methods in Comparative Safety and Effectiveness Research. 2008;In press.
137. Eng PM, Seeger JD, Loughlin J, Clifford CR, Mentor S, Walker AM. Supplementary data collection with case-cohort analysis to address potential confounding in a cohort study of thromboembolism in oral contraceptive initiators matched on claims-based propensity scores. Pharmacoepidemiol Drug Saf. 2008;17(3):297-305.
138. Sturmer T, Schneeweiss S, Avorn J, Glynn RJ. Correcting effect estimates for unmeasured confounding in cohort studies with validation studies using propensity score calibration. Am J Epidemiol. 2005;162:279-289.
139. Anand IS, Carson P, Galle E, Song R, Boehmer J, Ghali JK, Jaski B, Lindenfeld J, O'Connor C, Steinberg JS, Leigh J, Yong P, Kosorok MR, Feldman AM, DeMets D, Bristow MR. Cardiac resynchronization therapy reduces the risk of hospitalizations in patients with advanced heart failure: results from the Comparison of Medical Therapy, Pacing and Defibrillation in Heart Failure (COMPANION) trial. Circulation. 2009;119(7):969-977.
140. Rothman KJ. Modern Epidemiology 3rd Edition. Philadelphia: Lippincott Williams & Wilkins.
141. Schneeweiss S, Seeger JD, Maclure M, Wang PS, Avorn J, Glynn RJ. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. *Am J Epidemiol*. 2001;154(4):854-864.

142. Scheurer D, Choudhry N, Swanton KA, Matlin O, Shrank W. Association between different types of social support and medication adherence. *Am J Manag Care*. 2012;18(12):e461-467.

143. Zullig LL, Gellad WF, Moaddeb J, Crowley MJ, Shrank W, Granger BB, Granger CB, Trygstad T, Liu LZ, Bosworth HB. Improving diabetes medication adherence: successful, scalable interventions. *Patient preference and adherence*. 2015;9:139-149.

144. Cosentino F, Cannon CP, Cherney DZI, Masiukiewicz U, Pratley R, Dagogo-Jack S, Frederich R, Charbonnel B, Mancuso J, Shih WJ, Terra SG, Cater NB, Gantz I, McGuire DK, Investigators VC. Efficacy of Ertugliflozin on Heart Failure-Related Events in Patients With Type 2 Diabetes Mellitus and Established Atherosclerotic Cardiovascular Disease: Results of the VERTIS CV Trial. *Circulation*. 2020;142(23):2205-2215.

145. Feakins BG, McFadden EC, Farmer AJ, Stevens RJ. Standard and competing risk analysis of the effect of albuminuria on cardiovascular and cancer mortality in patients with type 2 diabetes mellitus. *Diagn Progn Res*. 2018;2:13.

146. Fine J, Gray R. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*. 1999;94:496-509.

147. Scheike TH, Zhang MJ. Flexible competing risks regression modeling and goodness-of-fit. *Lifetime Data Anal*. 2008;14(4):464-483.

148. Fralick M, Kim SC, Schneeweiss S, Everett BM, Glynn RJ, Patorno E. Risk of amputation with canagliflozin across categories of age and cardiovascular risk in three US nationwide databases: cohort study. *BMJ*. 2020;370:m2812.

149. Lesko CR, Henderson NC, Varadhan R. Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. *J Clin Epidemiol*. 2018;100:22-31.

150. Henderson NC, Louis TA, Wang C, Varadhan R. Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Serv Outcomes Res Methodol*. 2016;16(4):213-233.

151. Zhang Y, Laber EB, Tsiatis A, Davidian M. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*. 2015;71(4):895-904.

152. Garry EM, Buse JB, Gokhale M, Lund JL, Nielsen ME, Pate V, Sturmer T. Study design choices for evaluating the comparative safety of diabetes medications: An evaluation of pioglitazone use and risk of bladder cancer in older US adults with type-2 diabetes. *Diabetes Obes Metab*. 2019;21(9):2096-2106.

153. Lewis JD, Ferrara A, Peng T, Hedederson M, Bilker WB, Quesenberry CP, Jr., Vaughn DJ, Nessel L, Selby J, Strom BL. Risk of bladder cancer among diabetic patients treated with pioglitazone: interim report of a longitudinal cohort study. *Diabetes Care*. 2011;34(4):916-922.

154. Michels KB, Rosner BA. Data trawling: to fish or not to fish. *Lancet*. 1996;348(9035):1152-1153.

155. Orsini LS, Monz B, Mullins CD, Van Brun D, Daniel G, Eichler HG, Graff J, Guerino J, Berger M, Lederer NM, Jonsson P, Schneeweiss S, Wang SV, Crown W, Goettisch W, Willke RJ. Improving transparency to build trust in real-world secondary data studies for hypothesis testing: Why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Pharmacoepidemiol Drug Saf*. 2020.

156. Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, Klungel O, Petersen I, Sorensen HT, Dixon WG, Guttmann A, Harron K, Hemkens LG, Moher D, Schneeweiss S, Smee L, Sturkenboom M, von Elm E, Wang SV, Benchimol EI. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ*. 2018;363:k3532.
157. Wang SV. Reproducible Evidence: Practices to Enhance and Achieve Transparency (REPEAT). Vol 20202020.

158. Schneeweiss S, Brown JS, Bate A, Trifiro G, Bartels DB. Choosing Among Common Data Models for Real-World Data Analyses Fit for Making Decisions About the Effectiveness of Medical Products. *Clin Pharmacol Ther.* 2020;107(4):827-833.

159. Zhang Y, Thamer M, Kshirsagar O, Hernan M, A. CERBOT (Comparative Effectiveness Research Based on Observational Data to Emulate a Target Trial). Vol 20202020.

160. The International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practices (GPP). *Pharmacoepidemiology and Drug Safety.* 2008;17(2):200-208.

161. EMA. ENCePP Guide on Methodological Standards in Pharmacoepidemiology. 2014.

162. Gagne JJ, Han X, Hennessy S, Leonard CE, Chrischilles EA, Carnahan RM, Wang SV, Fuller C, Iyer A, Katcoff H, Woodworth TS, Archdeacon P, Meyer TE, Schneeweiss S, Toh S. Successful Comparison of US Food and Drug Administration Sentinel Analysis Tools to Traditional Approaches in Quantifying a Known Drug-Adverse Event Association. *Clin Pharmacol Ther.* 2016;100(5):558-564.

163. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med Care.* 2010;48(6 Suppl):S114-120.

164. Greenland S. For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. *European journal of epidemiology.* 2017;32(1):3-20.
Essential points

1) Real-world evidence (RWE) uses electronic data generated by the routine operation of the healthcare system to estimate the effectiveness of medical products in clinical practice.

2) RWE is increasingly considered in medical decision making and complements the findings of treatment efficacy that we derive from randomized trials.

3) Contemplating the target trial that we would like to do and then emulating this trial with real-world data (RWD) provides in study design clarity and helps avoid biases.

4) The underlying data need to be fit-for-purpose to answer the specific question at hand

5) The new-user, active-comparator cohort design has demonstrated its usefulness in RWE studies in diabetes, including the reproduction and prediction of randomized trial findings

6) Propensity score matching is a useful analytic approach to adjust for patient characteristics in the absence of baseline randomization

7) Transparency in the conduct and presentation of RWE studies is critical to allow reviewer to evaluate the study validity and have confidence in their decision making
Research data
- Data collected PRIMARILY for research
  - For purpose: Data specifically for study purpose
    - Examples: Framingham Study, Nurses’ Health Study, Cardiovasc Health Study, Slone Birth Defects Study, Some registries
  - Other purpose: Data intended for other studies
    - Examples: Framingham Study, Nurses’ Health Study, Some registries

Non-interventional data
- Data used SECONDARILY for research
  - Other purpose
    - Clinical documentation: EHR-based studies, NDI linkage, Lab test databases, Some registries
    - Administrative: Claims data studies, Geocoding/census, Vital statistics
    - Digital health devices: Fitness monitors, Glucose monitors, PRO apps, Some registries

Transaction data
- Other purpose

Typical RWE database studies
Accelerated Pathways

Initial Approval

Examples:
- Biomarker to clinical endpoint,
- efficacy claim expansion,
- broadened population

2nd Indication

Examples:
- a) Post-market requirements (PMR) or commitment (PMC),
- b) Rapid regulatory response to a safety signal

Primary approval

Examples:
- External control arms for single arm trials, RCT enrichment, RCT long-term follow-up

Secondary indications

Examples:
- Indication expansion, population expansion (pediatrics, stage), surrogate to clinical endpoint claim expansion

Accelerated Pathways

Initial Approval

Full Approval

Examples:
- Biomarker to clinical endpoint, efficacy claim expansion, broadened population

Secondary approval

Examples:
- External control arms for single arm trials, RCT enrichment, RCT long-term follow-up

Safety (a) and (b)
A transactional database records a stream of healthcare records in calendar time for all enrolled patients:

Stabilized data snapshot for research purposes and ensure reproducibility of findings

Individual patient-level data has arrived in batches from various sources and is arranged longitudinally

A causal study design is implemented in event time anchored at the cohort entry date
New-user active-comparator cohort study

Case-control sampling in the same cohort
**Achieve causal study design**
- Controlled or self-controlled
- Medically informed target population
- Patient-informed outcomes
- Biologically informed effect window

**Achieve fit-for-purpose measurement**
- Filling Rx
  - Prescribing Rx, self-report, infusers, pill caps, UDI from OR notes
- Dx, Px codes
  - Labs, imaging, digital health dev, physician notes, patient reports
- Dx, Px, Rx codes
  - Labs, stage, imaging, BMI, genomics, physician notes

**Achieve causal analysis**
- Achieve balance between groups:
  - Regression, PS analysis
  - Proxy adjustment: HDPS, CTMLE
  - Time-varying exposure: MSM
- Check balance:
  - Stand. difference, residuals, c-statistic

**BIAS REDUCTION**
- New users, active comparators
- Causal temporality
  - Exposure before outcome
  - Confounder before exposure

**DESIGN CHOICE**
- Controlled or self-controlled
- Medically informed target population
- Patient-informed outcomes
- Biologically informed effect window

**EXPOSURE**
- Sensitivity
- Specificity
- Completeness
- Mean sqr. diff.

**OUTCOME**
- Sensitivity
- Specificity
- Completeness
- Mean sqr. diff.

**CONFOUNDERS**
- Bio-monitors, physician notes, biomarker, omics, behavior, socioeconomic

**STUDY POPULATION**
- Bio-monitors, physician notes, biomarker, omics, behavior, socioeconomic
Does drug exposure vary within patients? no

Consider ...

Case-crossover study, SCCS
Randomized crossover trial

Does drug exposure vary between patients? no

Consider ...

Cohort study, CCS, CCoh, 2-SS
Randomized controlled trial

Does drug exposure vary between provider groups?

Consider ...

Instrumental variable analysis
Cluster randomized trial

SCCS: Self-Controlled Case Series
CCS: Case-Control Sampling
CCoh: Case-Cohort Sampling
2-SS: Two-Stage Sampling
Parallel group RCT

Washout period w/o study drug use

Exposed

Comparator (or placebo)

New-user, active-comparator cohort study

Washout period w/o study drug use

Exposed

Comparator

R = randomization; S = selection
Real-World Evidence Study

December 2018

Using Real-World Data to Predict Findings of an Ongoing Phase IV Cardiovascular Outcome Trial: Cardiovascular Safety of Linagliptin Versus Glimepiride

Randomized Controlled Trial

June 2019

Effect of Linagliptin vs Glimepiride on Major Adverse Cardiovascular Outcomes in Patients With Type 2 Diabetes

Risk of CV events (3P-MACE)

HR = 0.98 (0.84 – 1.14)

Risk of CV events (3P-MACE)

HR = 0.91 (0.79 – 1.05)
Diagnosis of T2DM

First drug exposure

Subsequent switch

Exposure classified as

SGLT-2i

SGLT-2i

If the oGLD user does not die before SGLT-2i
the oGLD person-time will not be counted

If the oGLD user does die before SGLT-2i
the oGLD person-time will be counted

Downloaded from https://academic.oup.com/endrv/advance-article/doi/10.1210/endrev/bnab007/6169054 by guest on 10 April 2021
Claims-defined patient characteristics:
120 pre-exposure variables in propensity score model:

| Claims-based patient characteristics | Other chronic ischemic heart disease, % | Coronary procedure (CABG or PTCA), % | History of PTCA or CABG, % | Ischemic stroke, % | Congestive heart failure, % | Renal dysfunction, % | Oedema, % | Use of medications, % | Past use of metformin | Past use of other antidiabetic agents | Past use of insulin | ACE inhibitor | ARBs | β-blockers | Thiazides | Loop diuretics | Calcium channel blockers | Statins | Other blood-lowering drugs | Oral anticoagulants | Antiplatelet | Healthcare utilization | Any hospitalization, % | Any hospitalization within prior 30 days, % | Mean (SD) hospital days | Mean (SD) number of physician visits | Mean (SD) distinct non-insulin antidiabetic prescriptions | Mean (SD) distinct prescriptions | Mean (SD) number of laboratory tests ordered |
|-------------------------------------|----------------------------------------|-------------------------------------|-----------------------------|----------------------|-------------------------|-------------------------|----------------|---------------------|------------------------|-----------------------------|--------------------------|----------------|------|--------|--------|----------------|--------------------------|----------------|----------------------|------------------------|----------------|----------------------|------------------------|------------------|-----------------|------------------|
| Demographics                        | Mean (SD) age                          | Female, %                           | Features of medication initiation, % | Monotherapy          | Dual therapy            | Therapy with >2 agents | Dual therapy with metformin | Concomitant initiation of other antidiabetic agents, % | Concomitant initiation of insulin | Current use of other antidiabetic agents, % | Current use of metformin | Current use of insulin |
|                                      |                                        |                                     |                             |                      |                        |                          |                            |                                  |                          |                            |                          |                          |
| Comorbidities at baseline, %        | Mean (SD) Charlson comorbidity score   | Diabetic nephropathy, %             | Diabetic retinopathy, %     | Diabetic neuropathy, % | Peripheral vascular disease, % | Extrinsic dyspnoea, % | Diabetic foot, % | Skin infections, % | Hypoglycemia, % | Hyperparathyroidism, % | Hyperlipidaemia, % | Coronary atherosclerosis, % | Acute myocardial infarction, % | Old myocardial infarction, % | Unstable angina, % | Stable angina, % |
|                                      |                                        |                                     |                             |                      |                        |                          |                            |                                  |                          |                            |                          |                          |

EHR-defined patient characteristics:
Balance of 6 characteristics of specific concern for causal inference:

- Smoking status
- BMI
- Duration of DM
- HbA1c level
- GFR value
- LDL cholesterol level

RWE study of linagliptin vs. pioglitazone and cardiovascular endpoints (MACE)

- 18,804 T2DM patients in claims data initiating linagliptin vs. pioglitazone
- 7,926 T2DM patients after 1:1 propensity score matching
- 7,219 patients had linkable EHR data available
- 1,159 patients with EHR data after 1:1 propensity score matching

| Characteristics                | Linagliptin | Pioglitazone |
|---------------------------------|-------------|--------------|
| Never smoking                   | 32.4%       | 33.9%        |
| Obese                           | 49.4%       | 46.1%        |
| >3 years DM duration            | 17.7%       | 20.1%        |
| HbA1C, %                        | 8.0 (7.1-9.1) | 8.2 (7.1-9.9) |
| eGFR, ml/min/1.73m²             | 102 (93-116) | 104 (96-118) |
| LDL, mg/dl                      | 97 (73-116)  | 97 (79-115)  |
Balancing patient characteristics

- Define pre-exposure covariates based on clinical knowledge
- Define additional covariates empirically (high-dimensional proxy adjustment)
- Demonstrate covariate balance by treatment group: RDs and standardized differences
- Supplemental covariate information required that is not available in primary data source?
  - Yes
  - Collect additional information in sub-population
    - 2-stage sampling
    - External data source (PS Calibration)
    - Multiple imputation

Propensity score analysis

- Missing covariate values in EHRs?
  - Yes
  - Consider multiple imputation
  - Consider continuity cohorts
- Estimate propensity score (PS)
- Graphically explore PS distribution by treatment group
- Explore effect measure modification by PS: tabulate RR, RD for each PS stratum
- Effect measure modification by PS?
  - Yes
  - Match on PS (1:1, 1:n, n:m)
  - Stratify by PS deciles
- Demonstrate covariate balance by treatment group: RDs and standardized differences

Trim 5% of patients on each end of PS distribution or match by PS
Repeat analyses after changes in:
• Definition of "new users"
• Definition of exposure risk window
• Outcome definition if appropriate

Explore changes in effect estimates after making structural assumptions about unmeasured confounders

Calculate risk difference (RD) and risk ratio (RR) with 95% confidence intervals (CIs);
Person-time (p-t), number of events

Subgroup analysis
Calculate RR, RD for each subgroup

Dose-response analysis

Statistical analysis

Sensitivity analyses

Report