Received: 17 September 2019 | Revised: 29 October 2019 | Accepted: 9 November 2019

DOI: 10.1111/dom.13918

WILEY

REVIEW ARTICLE

Transparency in real-world evidence (RWE) studies to build confidence for decision-making: Reporting RWE research in diabetes

Elisabetta Patorno MD | **Sebastian Schneeweiss MD** | **Shirley V. Wang PhD**

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

Correspondence
Elisabetta Patorno, MD, Dr PH, Assistant Professor of Medicine, 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, USA.
Email: epatorno@bwh.harvard.edu.

Funding information
This study was funded by the Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts. E.P. was supported by a career development grant (K08AG055670) from the National Institute on Aging.

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1111/dom.13918.

Abstract

Transparency of real-world evidence (RWE) studies is critical to understanding how findings of a specific study were derived and is a necessary foundation to assessing validity and determination of whether decisions should be informed by the findings. In the present paper, we lay out strategies to improve clarity in the reporting of comparative effectiveness studies using real-world data that were generated by the routine operation of a healthcare system. This may include claims data, electronic health records, wearable devices, patient-reported outcomes or patient registries. These recommendations were discussed with multiple stakeholders, including regulators, payers, academics and journal editors, and endorsed by two professional societies that focus on RWE. We remind readers interested in diabetes research of the utility of conceptualizing a target trial that is then emulated by a RWE study when planning and communicating about RWE study implementation. We recommend the use of a graphical representation showcasing temporality of key longitudinal study design choices. We highlight study elements that should be reported to provide the clarity necessary to make a study reproducible. Finally, we suggest registering study protocols to increase process transparency. With these tools the readership of diabetes RWE studies will be able to more efficiently understand each study and be more able to assess a study's validity with reasonably high confidence before making decisions based on its findings.

KEYWORDS
cohort study, pharmaco-epidemiology, population study, type 2 diabetes

1 | INTRODUCTION

Recent cardiovascular outcome trials of sodium-glucose co-transporter-2 (SGLT2) inhibitors to treat type 2 diabetes have demonstrated substantial reduction in hospitalization for heart failure and cardiovascular events.[1] These findings have now been replicated in non-interventional database studies that make use of "real-world" data (RWD) to generate real-world evidence (RWE)[2,3]; however, would we have believed the RWE studies in the absence of the findings from randomized controlled trials (RCTs)? Why is it that we have so much more confidence in RCTs than in RWE studies? Skepticism of RWE is justifiable. There are plenty of examples where RWE studies were in complete contradiction of RCTs. Think of hormone replacement therapy (HRT) in post-menopausal women, where a postulated reduction in risk of coronary heart disease was later found to increase the initial risk; vitamin E supplementation was thought to be protective of coronary heart disease but the effects could never be reproduced in a trial, and the substantial reduction in fractures and dementia associated with statin use in RWE studies were not borne out by RCTs.[4-6]

As pharmaco-epidemiologists, we need to acknowledge that there has been and will continue to be publication of misleading non-interventional RWE; however, we have made enormous progress over the past 20 years and, for the most part, we now understand how avoidable biases in design or analysis led prior RWE studies to miss the mark.[7,8] Re-analyses of HRT cohort studies have shown that using a new-user design mimicking a parallel group trial instead of a current-user analysis could completely correct the operating bias.[9] The unrealistically large mortality benefit reported by some RWE studies of SGLT2 inhibitors,[1,10,11] but not by others could be explained by immortal-time bias, a bias which can be avoided with appropriate attribution of person-time for the compared exposures.[2,3]

The first step to understanding the validity of RWE studies is to understand what exactly was done in a given study. How were the data curated and what were the transformations performed on 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, in a way which is understandable to most. 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.[12]

For RWE to have maximum impact, 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 RCTs, we need to provide decision-makers with unambiguous reporting of RWE study conduct, provide tools to facilitate efficient review, and provide guidance on how to assess the validity of results.[13] 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.

Because RWE makes secondary use of existing streams of healthcare data, theoretically, independent investigators should be able to implement the same protocols in the same data source and obtain the exact same results independently. However, recent efforts to replicate studies have found that reporting about the methods used to generate RWE, including specific code algorithms, temporality of assessing exposures, inclusion criteria, covariates and outcome is often too ambiguous for independent teams to closely replicate published findings.[14-19]

In the following sections, we lay out strategies to improve clarity in the reporting of comparative effectiveness studies using RWD that were generated by the routine operation of a healthcare system. This may include claims data, electronic health records, wearable devices, patient-reported outcomes, or patient registries.[20] These recommendations were discussed with multiple stakeholders, including, regulators, payers, academics and journal editors, and endorsed by two professional societies that focus on RWE, the International Society of Pharmacoepidemiology and the International Society for Pharmacoeconomics and Outcomes Research,[13,21,22] and aim to provide guidance to the readership of RWE diabetes studies to understand each study more efficiently and assess a study's validity before making decisions based on its findings.

2 | CONCEPTUALIZING A TARGET TRIAL THAT THE RWE IS TRYING TO EMULATE

There are many reasons why 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 because they both try to establish causal relationships between medical products or interventions and health outcomes. Before intervening with patients, we want to ensure that we are treating them with medical products that improve health outcomes.

It has therefore been proposed multiple times over decades, and most specifically by Hernán and Robins,[23] that by envisioning the design of 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 RWD 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 temporality, that is, determining when patient characteristics, exposure and outcomes are measured relative to study entry, which is critical to enabling causal conclusions to be drawn. It clarifies the analytical strategy of an intention-to-treat analysis or an on-treatment analysis. Once a target trial is conceptualized, the design of the trial-emulating RWE study and potential diversion from the trial makes clear potential weaknesses in data quality, data completeness, and causal inference. It is hoped that such clarity will lead to adjustments in the RWE study that improve validity.[24] A trial-emulating RWE study design often exposes a tension between the objective to have highly generalizable findings in RWE studies and the restrictions that need to be imposed to ensure high validity of the findings that will allow causal conclusions.

We see the target trial approach as a meaningful quality improvement strategy when planning and when communicating about RWE studies. Our recommendations on unambiguous reporting of RWE studies follow this paradigm.

3 | CLARITY REGARDING BASIC TEMPORALITY OF LONGITUDINAL DESIGN CHOICES

Real-world evidence studies make secondary use of non-interventional data that were not collected for research purposes.[20] Thus, they often involve complex design and analysis decisions. It is vital to enable readers and decision-makers to understand quickly yet comprehensively the basic temporality of the study design used to generate RWE. A group of experts and advisors from academia, regulatory, publishers, payers and industry, therefore, proposed a visualization schematic that illustrates comparative effectiveness

study designs with longitudinal data.[22] In this section, we summarize the proposed visualization framework.

Because of the complexity of the timeline and the inter-related nature of the factors described above, researchers often find it helpful to illustrate their study design implementation on an imaginary patient longitudinal healthcare record; however, if a design diagram is presented, the design elements represented in the diagram in published reports varies widely.[25-29] We proposed a framework for visualizing study design using standardized structure and terminology. The framework focuses on summarizing details of first- and second-order temporal anchors (Table 1). First-order anchors are represented by columns and second-order anchors, which are defined relative to first-order anchors, are visually defined by horizontal boxes (Figure 1). In addition to boxes that visually represent temporality relative to the first-order anchor of cohort entry date (day 0), the design diagrams include bracketed numbers representing inclusive time intervals (following conventional mathematical notation). These diagrams are designed to be read from top to bottom, indicating the steps taken to create an analytical study population. The diagrams could be enhanced by inclusion of patient counts, showing the flow of patients that might typically be found in an attrition table or CONSORT diagram, at each sequential box. We provide one example of a basic cohort study comparing risk of angioedema for angiotensin-converting enzyme inhibitors versus angiotensin receptor blockers.[30] Additional examples of diagrams for different study designs, including cohort designs, cohort sampling designs (case–control, case-cohort, two-stage sampling) and self-controlled designs can be found at: https://www.repeatinitiative.org/projects.html

4 | UNAMBIGUOUS REPORTING OF DATA TRANSFORMATIONS TO PREPARE STUDY DATA

In this section, we summarize a catalogue of specific design and implementation parameters outlined by a joint task force between the International Society for Pharmacoepidemiology (ISPE) and the International Society for Pharmacoeconomics and Outcomes Research (ISPOR).[13] 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.

4.1 | Source data characteristics

Researchers should describe characteristics of the source data, including specifying the data extraction date or data version and range in years of source data available (Table 2A). Data may have subtle or profound differences depending on when the raw source data are cut by the data provider for research use, so even if an investigator uses the same code on data from the same data source, they may obtain different results if the source data are cut at different time points.

TABLE 1 Temporal anchors in a longitudinal study of drug effects

Base anchor (defined in calendar time, describes source data)		
DED	**Data extraction date**	**The date when the data were extracted from the dynamic transactional database**
SDR	Source data range	The calendar time range covered by a data source that is available to create the study population.
SP	Study period	The calendar time boundaries for data used to create the analyzed study dataset including exposures, inclusion/exclusion criteria, covariates, outcome and follow-up.
First-order anchor (defined in patient event time, specifies study entry/index date)		
SED	Study entry date	The date when subjects enter the study population.
OED	Outcome event date[a]	The date of an outcome event occurrence.
Second-order anchors (defined in patient event time, relative to first-order anchor)		
W_E	Washout window for exposure	An interval used to define incident exposure. If there is no record of exposure (and/or comparator) of interest within this interval, the next exposure is considered "new" initiation, otherwise it is considered prevalent exposure.
W_O	Washout window for outcome	An interval used to define incident outcomes. If there is no record of outcomes within this interval, the next outcome is considered incident.
EXCL	Exclusion assessment window	An interval during which patient exclusion criteria are assessed.
CAW	Covariate assessment window	An interval during which patient covariates are assessed. The CAW should precede the EAW in order to avoid adjusting for causal intermediates. It is sometimes called baseline period.
EAW	Exposure assessment window	The window during which exposure status is assessed. The exposure status is defined at the end of the EAW.[b] The EAW should precede the FUW to avoid reverse causation.
FUW	Follow-up window	The interval during which occurrence of the outcome of interest in the study population will be included in the analysis. The FUW may involve stockpiling algorithms, grace periods, exposure extension and/or censoring related to exposure discontinuation.

[a]The outcome event date can be a first-order anchor in some study designs (eg, case-crossover, case–control).
[b]This is relevant in sampling designs when the occurrence of exposure is not a first-order anchor defining cohort entry.

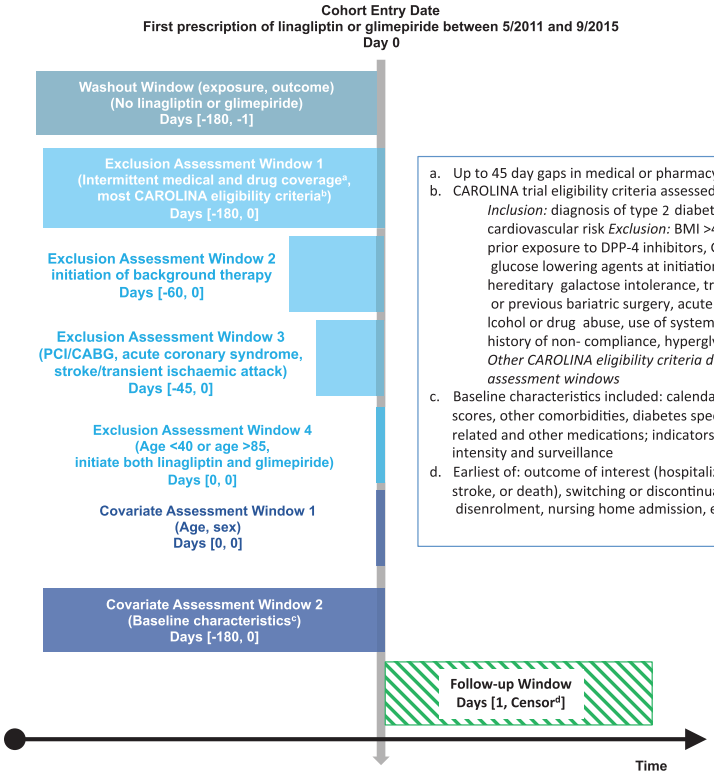

FIGURE 1 Example Design Diagram: the CAROLINA trial prediction with real-world evidence.[53] BMI, body mass index; CABG, coronary artery bypass grafting; GLP-1, glucagon-like peptide-1; PCI, percutaneous coronary intervention; SGLT2, sodium-glucose co-transporter-2; TZD, thiazolidinediones

Similarly, researchers should describe the types of data available in the data source. Are the data based on insurance claims, electronic health records, disease registries, or other sources of data? Are there *de novo* data linkages? Which subsets of data from which sources were accessible to the investigators? The sampling strategy and any inclusions or exclusions applied to obtain a cut of source data should be reported. For example, Medicare claims data in the United States can be provided as a random 5% sample or based on tailored investigator selection criteria (eg, presence of an inpatient or outpatient diabetes diagnosis in the years 2012–2018). This type of information about the source data, based upon which investigators create their analytical cohorts, will help readers understand the implications regarding completeness of data capture and missingness as well as interpretation of the validity of findings.

When the raw data provided by a vendor are pre-processed by the investigative team, before creating an analytical cohort for a study, this process should be described. For example, cleaning "messy" data fields or imputing missing data on a database-wide level, or for a specific project. Sometimes raw data are converted to a common data model. When that is the case, the common data model version should be referenced. For example, one might state that the data were converted to fit the US Food and Drug Administration Sentinel Common Data Model version 7.0.0. Materials detailing any assumptions applied during the data conversion process, as well as dates of refreshes if the stored data are periodically updated with more recent data, should also be made available as citable resources.[31,32]

4.2 | Cohort entry criteria, exposure, outcome, follow-up and covariates

When describing an analytical study population, it is not sufficient to simply state the names of the inclusion/exclusion criteria, exposures, outcomes and covariates being investigated. There are several layers of detail necessary to fully define these measures. Reporting the specific codes used to define these measures is critical for clear communication of study methodology, to the point of reproducibility,[19,33] particularly in databases where there may be substantial ambiguity and investigator discretion in code choice (eg, READ codes in UK data). Other key elements that should be unambiguously reported include details about diagnosis positions and care settings in which to identify the relevant codes, criteria to ensure capture of patient healthcare contacts in the source data (eg, enrolment, up-to-standard research data), and temporality of measurement relative to cohort entry (Table 2C).

It is critical to provide detailed description of criteria that define who is included in a study. In addition to details about specific inclusion/exclusion measures (codes, temporality, care setting, diagnosis position), other key operational decisions to communicate include whether patients are allowed to enter the cohort once or multiple times, clarity about the cohort entry defining criterion and whether the date of cohort entry is selected before or after application of other exclusion criteria. These implementation decisions determine which patients are included and when they enter the cohort; different decisions could result in different person-time from the same patients contributing to the analysis.

TABLE 2 Reporting specific parameters to increase reproducibility of database studies

	Description	Example	Synonyms
A. Reporting on data source should include:			
A.1 Data provider	Data source name and name of organization that provided data.	Medicaid Analytic Extracts data covering 50 states from the Centres for Medicare and Medicaid Services.	
A.2 DED	The date (or version number) when data were extracted from the dynamic raw transactional data stream (eg, date that the data were cut for research use by the vendor).	The source data for this research study was cut by [data vendor] on January 1, 2017. The study included administrative claims from January 1, 2005 to December 31, 2015.	Data version, data pull
A.3 Data sampling	The search/extraction criteria applied if the source data accessible to the researcher is a subset of the data available from the vendor.		
A.4 SDR	The calendar time range of data used for the study. Note that the implemented study may use only a subset of the available data.		Study period, query period
A.5 Type of data	The domains of information available in the source data, eg, administrative, electronic health records, inpatient vs. outpatient capture, primary vs. secondary care, pharmacy, laboratory, registry.	The administrative claims data include enrolment information, inpatient and outpatient diagnosis (ICD-9/−10) and procedure (ICD-9/−10, CPT, HCPCS) codes as well as outpatient dispensations (NDC codes) for 60 million lives covered by Insurance X. The electronic health records data include diagnosis and procedure codes from billing records, problem list entries, vital signs, prescription and laboratory orders, laboratory results, inpatient medication dispensation, as well as unstructured text found in clinical notes and reports for 100 000 patients with encounters at ABC integrated healthcare system.	
A.6 Data linkage, other supplemental data	Data linkage or supplemental data such as chart reviews or survey data not typically available with licence for healthcare database.	We used SEER data on prostate cancer cases from 1990 through 2013 linked to Medicare and a 5% sample of Medicare enrollees living in the same regions as the identified cases of prostate cancer over the same period of time. The linkage was created through a collaborative effort from the NCI, and the CMS.	
A.7 Data cleaning	Transformations to the data fields to handle missing, out of range values or logical inconsistencies. This may be at the data source level or the decisions can be made on a project-specific basis.	Global cleaning: The data source was cleaned to exclude all individuals who had more than one gender reported. All dispensing claims that were missing day's supply or had 0 days' supply were removed from the source data tables. Project-specific cleaning: When calculating duration of exposure for our study population, we ignored dispensation claims that were missing or had 0 days' supply. We used the most recently reported birth date if there was more than one birth date reported.	
A.8 Data model conversion	Format of the data, including description of decisions used to convert data to fit a CDM.	The source data were converted to fit the Sentinel CDM version 5.0. Data conversion decisions can be found on our website (http://ourwebsite). Observations with missing or out of range values were not removed from the CDM tables.	
B. Reporting on overall design should include:			
B.1 Design diagram	A figure that contains 1st and 2nd order temporal anchors and depicts their relation to each other.	See example in Figure 1.	

(Continues)

TABLE 2 (Continued)

	Description	Example	Synonyms
C. Reporting on inclusion/exclusion criteria should include:			
C.1 SED	The date(s) when subjects enter the cohort.	We identified the first SED for each patient. Patients were included if all other inclusion/exclusion criteria were met at the first SED.	Index date, cohort entry date, outcome date, case date, qualifying event date, sentinel event
C.2 Person or episode level study entry	The type of entry to the cohort. For example, at the individual level (1× entry only) or at the episode level (multiple entries, each time inclusion/exclusion criteria met).	We identified all SEDs for each patient. Patients entered the cohort only once, at the first SED where all other inclusion/exclusion criteria were met.	Single vs. multiple entry, treatment episodes, drug eras
C.3 Sequencing of exclusions	The order in which exclusion criteria are applied, specifically whether they are applied before or after the selection of the SED(s).	We identified all SED for each patient. Patients entered the cohort at every SED where all other inclusion/exclusion criteria were met.	Attrition table, flow diagram, CONSORT diagram
C.4 EW	The time window prior to SED in which an individual was required to be contributing to the data source.	Patients entered the cohort on the date of their first dispensation for Drug X or Drug Y after at least 180 days of continuous enrolment (30 day gaps allowed) without dispensings for either Drug X or Drug Y.	Observation window
C.5 Enrolment gap	The algorithm for evaluating enrolment prior to SED including whether gaps were allowed.		
C.6 Inclusion/Exclusion definition window	The time window(s) over which inclusion/exclusion criteria are defined.	Exclude from cohort if ICD-9 codes for deep vein thrombosis (451.1x, 451.2x, 451.81, 451.9x, 453.1x, 453.2x, 453.8x, 453.9x, 453.40, 453.41, 453.42 where x represents presence of a numeric digit 0–9 or no additional digits) were recorded in the primary diagnosis position during an inpatient stay within the 30 days prior to and including the SED. Invalid ICD-9 codes that matched the wildcard criteria were excluded.	
C.7 Codes	The exact drug, diagnosis, procedure, laboratory or other codes used to define inclusion/exclusion criteria.		Concepts, vocabulary, class, domain
C.8 Frequency and temporality of codes	The temporal relation of codes in relation to each other as well as the SED. When defining temporality, be clear whether or not the SED is included in assessment windows (eg, occurred on the same day, two codes for A occurred within 7 days of each other during the 30 days prior to and including the SED).		
C.9 Diagnosis position (if relevant/available)	The restrictions on codes to certain positions, eg, primary vs. secondary. Diagnoses.		
C.10 Care setting	The restrictions on codes to those identified from certain settings, eg, inpatient, emergency department, nursing home.		Care site, place of service, point of service, provider type
C.11 Washout for exposure	The period used to assess whether exposure at the end of the period represents new exposure.		Lookback for exposure, event-free period
C.12 Washout for outcome	The period prior to SED or OED to assess whether an outcome is incident.	Patients were excluded if they had a stroke within 180 days prior to and including the cohort entry date.Cases of stroke were excluded if there was a recorded stroke within 180 days prior.	Lookback for outcome, event-free period

(Continues)

TABLE 2 (Continued)

	Description	Example	Synonyms
D. Reporting on exposure definition should include:			
D.1 Type of exposure	The type of exposure that is captured or measured, eg, drug versus procedure, new use, incident, prevalent, cumulative, time-varying.	We evaluated risk of outcome Z following incident exposure to drug X or drug Y. Incident exposure was defined as beginning on the day of the first dispensation for one of these drugs after at least 180 days without dispensations for either (SED). Patients with incident exposure to both drug X and drug Y on the same SED were excluded. The exposure risk window for patients with Drug X and Drug Y began 10 days after incident exposure and continued until 14 days past the last days supply, including refills. If a patient refilled early, the date of the early refill and subsequent refills were adjusted so that the full days supply from the initial dispensation was counted before the days supply from the next dispensation was tallied. Gaps of ≤14 days in between one dispensation plus days supply and the next dispensation for the same drug were bridged (ie, the time was counted as continuously exposed). If patients exposed to Drug X were dispensed Drug Y or vice versa, exposure was censored. NDC codes used to define incident exposure to drug X and drug Y can be found in the appendix.	
D.2 ERW	The ERW is specific to an exposure and the outcome under investigation. For drug exposures, it is equivalent to the time between the minimum and maximum hypothesized induction time following ingestion of the molecule.		Drug era, risk window
D.2a Induction period	Days on or following study entry date during which an outcome would not be counted as "exposed time" or "comparator time".		Blackout period
D.2b Stockpiling	The algorithm applied to handle leftover days supply if there are early refills.		
D.2c Bridging exposure episodes	The algorithm applied to handle gaps that are longer than expected if there was perfect adherence (eg, non-overlapping dispensation + day's supply).		Episode gap, grace period, persistence window, gap days
D.2d Exposure extension	The algorithm applied to extend exposure past the days supply for the last observed dispensation in a treatment episode.		Event extension
D.3 Switching/add-on	The algorithm applied to determine whether exposure should continue if another exposure begins.		Treatment episode truncation indicator
D.4 Codes, frequency and temporality of codes, diagnosis position, care setting	Description in Section C.	Drug X was defined by NDC codes listed in the appendix. Brand and generic versions were used to define Drug X. Non pill or tablet formulations and combination pills were excluded.	Concepts, vocabulary, class, domain, care site, place of service, point of service, provider type

(Continues)

TABLE 2 (Continued)

	Description	Example	Synonyms
D.5 EAW	A time window during which the exposure status is assessed. Exposure is defined at the end of the period. If the occurrence of exposure defines cohort entry, eg, new initiator, then the EAW may be a point in time rather than a period. If EAW is after cohort entry, FUW must begin after EAW.	We evaluated the effect of treatment intensification vs. no intensification following hospitalization on disease progression. Study entry was defined by the discharge date from the hospital. The exposure assessment window started from the day after study entry and continued for 30 days. During this period, we identified whether or not treatment intensified for each patient. Intensification during this 30-day period determined exposure status during follow-up. Follow-up for disease progression began 31 days following study entry and continued until the first censoring criterion was met.	
E. Reporting on follow-up time should include:			
E.1 FUW	The time following cohort entry during which patients are at risk to develop the outcome due to the exposure. FUW is based on a biologic exposure risk window defined by minimum and maximum induction times. However, FUW also accounts for censoring mechanisms.	Follow-up began on the SED and continued until the earliest of discontinuation of study exposure, switching/adding comparator exposure, entry to nursing home, death, or end of study period.	
E.2 Censoring criteria	The criteria that censor follow-up.	We included a biologically plausible induction period, therefore, follow-up began 60 days after the SED and continued until the earliest of discontinuation of study exposure, switching/adding comparator exposure, entry to nursing home, death, or end of study period.	
F. Reporting on outcome definition should include:			
F.1 OED	The date of an event occurrence.	The OED was defined as the date of first inpatient admission with primary diagnosis 410.x1 after the SED and occurring within the follow-up window.	Case date, measure date, observation date
F.2 Codes, frequency and temporality of codes, diagnosis position, care setting	Description in Section C.		Concepts, vocabulary, class, domain, care site, place of service, point of service, provider type
F.3. Validation	The performance characteristics of outcome algorithm if previously validated.	The outcome algorithm was validated via chart review in a population of diabetics from data source D (citation). The positive predictive value of the algorithm was 94%.	
G. Reporting on covariate definitions should include:			Event measures, observations
G.1 CAW	The time over which patient covariates are assessed.	We assessed covariates during the 180 days prior to but not including the SED.	Baseline period
G.2 Comorbidity/ risk score	The components and weights used in calculation of a risk score.	See appendix for example. Note that codes, temporality, diagnosis position and care setting should be specified for each component when applicable.	
G.3 Healthcare utilization metrics	The counts of encounters or orders over a specified time period, sometimes stratified by care setting, or type of encounter/order.	We counted the number of generics dispensed for each patient in the CAW. We counted the number of dispensations for each patient in the CAW. We counted the number of outpatient encounters recorded in the CAW. We counted the number of days with outpatient encounters recorded in the CAW. We counted the number of inpatient hospitalizations in the CAW, if admission and discharge dates for different encounters overlapped, these were "rolled up" and counted as 1 hospitalization.	

(Continues)

TABLE 2 (Continued)

	Description	Example	Synonyms
G.4 Codes, frequency and temporality of codes, diagnosis position, care setting	Description in Section C.	Baseline covariates were defined by codes from claims with service dates within 180 days prior to and including the SED. Major upper gastrointestinal bleeding was defined as inpatient hospitalization with: At least one of the following ICD-9 diagnoses: 531.0x, 531.2x, 531.4x, 531.6x, 532.0x, 532.2x, 532.4x, 532.6x, 533.0x, 533.2x, 533.4x, 533.6x, 534.0x, 534.2x, 534.4x, 534.6x, 578.0- OR -An ICD-9 procedure code of: 44.43 - OR -A CPT code 43255	Concepts, vocabulary, class, domain, care site, place of service, point of service, provider type
H. Reporting on control sampling should include:			
H.1 Sampling strategy	The strategy applied to sample controls for identified cases (patients with OED meeting all inclusion/exclusion criteria).	We used risk set sampling without replacement to identify controls from our cohort of patients with diagnosed diabetes (inpatient or outpatient ICD-9 diagnoses of 250.xx in any position). Up to 4 controls were randomly matched to each case on length of time since SED (in months), year of birth and gender. The random seed and sampling code can be found in the online appendix.	
H.2 Matching factors	The characteristics used to match controls to cases.		
H.3 Matching ratio	The number of controls matched to cases (fixed or variable ratio).		
I. Reporting on statistical software should include:			
I.1 Statistical software program used	The software package, version, settings, packages or analytic procedures.	We used: SAS 9.4 PROC LOGISTICCran R v3.2.1 survival packageSentinel's Routine Querying System version 2.1.1 CIDA+PSM[1] tool Aetion Platform release 2.1.2 Cohort Safety	

Abbreviations: CDM, Common Data Model; CMS, Centres for Medicare and Medicaid Services; CPT, current procedural terminology; CAW, covariate assessment window; DED, data extraction date; EAW, exposure assessment window; OED, outcome event date; ERW, exposure risk window; EW, enrolment window; FUW, follow-up window; HCPCS, healthcare common procedure coding system; ICD, International Classification of Diseases; NCI, National Cancer Institute; NDC, national drug code; SDR, source data range; SED, study entry date; SEER, Surveillance, Epidemiology, and End Results; Parameters in boldface are key temporal anchors.

Reporting about exposure should include description of the type of exposure being investigated, for example, new users (or incident users), current (or prevalent) users, or a mix of both.[34] When a study is investigating new users, the criteria used to define incident users should be clearly specified, including reporting what exposures patients are required to be naïve to (eg, drug of interest only, entire drug class, both exposure and comparator drugs), and the duration of the washout period to define incident users.

In addition to being clear about algorithms used to define the start of exposure, it is important to provide detail about how duration of exposure is defined. Duration is operationally defined based on the investigators decisions about how to handle recorded information about prescription or dispensed amounts and days' supply, observed early refills or gaps in between dispensations, and the hypothesized half-life for the effect of exposure. Algorithms can be applied to bridge observed exposure episodes and extend the hypothesized risk window beyond the end of an observed days' supply to allow for modest non-adherence as well as biological exposure risk window. These decisions can influence which days patients are counted as being at risk while exposed and which outcomes are counted in the analysis. (Table 2D).

Obviously, which outcomes are included in the analysis can be very influential when it comes to study findings. When interested in studying incident outcomes, the investigator may specify a minimum washout period during which there are no outcome codes prior to the cohort entry date or prior to the index outcome occurrence date. In addition to being clear about defining the outcome measure (eg, codes, care setting, diagnosis position), outcome ascertainment requires delineating temporality. The temporality for outcome ascertainment can be affected by censoring mechanisms other than operational decisions used to measure exposure duration, such as death, disenrolment, entry to a nursing home, add-on or switching of medications, or use of a fixed follow-up window (intention to treat). The algorithms to determine start and end of outcome surveillance should be communicated clearly so the reader can understand how days at risk are defined

and recognize potential biases that may arise with some choices (eg, informative censoring; Table 2E).

Similar to inclusion/exclusion criteria, exposure, and outcome measures, when a comorbidity score is used for a cohort, each component of the score should be clearly defined in terms of codes, care setting, diagnosis position and temporality. The weights for components should also be specified (Table 2F). This information can be contained in cited material, but papers often report evaluation of multiple versions of a score, so investigators need to make sure that they are clear about which version is used in their analysis when citing such papers. Equal clarity should be provided for healthcare utilization metrics, specifically, how the metric is calculated. For example, utilization can be considered unique by encounter or by day. Metrics may consider all care settings or only a specific subset (eg, inpatient). Different choices will result in different counts.

For cohort sampling designs, such as nested case–control studies, investigators should clearly communicate how and when controls are sampled from the source population (Table 2G). This could be base case, risk set, or survivor sampling. In addition to how controls are sampled, investigators should provide details on matching factors, what they are, how they are defined, the matching ratio and whether the ratio is fixed or variable.

The items described in this section and detailed in Table 1 are important to communicate unambiguously what was done to generate evidence for a RWE study and make the findings reproducible. High-level key temporal anchors should be reported with the design diagram in the methods sections of a paper. Given word count limits, supporting details may be provided in online appendices.

4.3 | Descriptive and comparative results

If patient counts are not incorporated into a design diagram, an attrition table that reports patient counts after implementing each inclusion/exclusion criterion should be reported for every RWE study conducted using large healthcare databases (Table 3A). Descriptive tables of the study population should include exposure-stratified columns describing the number of patients, baseline patient characteristics, person-years of follow-up, censoring reasons, number of health outcomes of interest, and measures of occurrence such as risks and rates. These descriptive tables characterize the cohort and facilitate assessment of whether a reproduction effort was successful. For comparative studies, measures of how comparable patients in the compared groups are should be provided. For evaluations of drugs or other medical products, the comparison would be across levels of exposure. For instrumental variable analysis, characteristics would be compared across levels of the instrument.[35]

Metrics for comparability across groups could include absolute or standardized differences for individual baseline characteristics or summary measures of differences such as the Mahalanobis distance (Table 3B).[36]

4.4 | Comparative analysis methods

Regardless of analytical method used, unadjusted as well as adjusted results should always be reported for both relative (hazard ratio, rate ratio, risk ratio, odds ratio) as well as absolute measures of association (rate difference, risk difference). In addition, researchers should be explicit about what quantity is being estimated, an average treatment effect versus average treatment effect among the treated, a marginal versus conditional effect, an intention-to-treat versus an on-treatment analysis, or other (Table 3C). In addition, researchers should be clear about which variables are used for adjustment, how they are parameterized, and how standard errors are obtained (eg, model-based, robust, bootstrap). We outline additional reporting expectations for some of the most commonly used cohort analysis methods, however, we recognize that there are alternative analysis methods that are not covered.

When only a small number of covariates are adjusted for, such as age and sex, stratification and standardization methods can be used.[37] If direct or indirect standardization is used, the standard (reference) population should be clearly defined. The covariates used and how they were categorized should be reported when either standardization or stratification methods are used to adjust for confounding. If there are more than a few covariates to adjust for, multivariable outcome regression can be used. When a multivariable outcome model is used, all coefficients from the model should be reported.

Propensity scores are another way of adjusting for numerous confounders. If a propensity score is used to summarize confounders into a single scalar, a propensity score distribution plot should be provided to show the range in score and degree of overlap for the exposure groups. Measures of the predictive accuracy of the strategy used to estimate the propensity score, for example, the c-statistic in a logistic regression model should also be provided. When propensity scores are used to match patients across levels of exposure, matched and unmatched tables of baseline characteristics with comparability metrics described above should be presented. Similarly, stratified tables and weighted tables of baseline characteristics should be presented when conducting propensity score-stratified or weighted analyses.

When propensity score matching is used to adjust for confounding, the algorithm for matching (eg, nearest-neighbour, greedy, full), caliper (eg, 0.2 standard deviations of the propensity score on the logit scale) and the matching ratio (eg, fixed 1:1, variable 1:4) should be reported. For studies that use one-to-one matching, the data can be validly analysed unconditionally (ignoring the matching) or conditionally (taking matching into account) and investigators should be clear how they analysed the data.[38] When patients are matched on factors other than propensity score, investigators should describe how the pool of potential matches was defined and the parameters used for matching. These operational specifications can influence results and therefore should be reported as part of study methods.

When confounding adjustment is based on stratification or weighting with the propensity score, investigators should be clear

TABLE 3 Reporting of descriptive and comparative results

	Description
A. Reporting of descriptive results should include:	
Flow diagram/attrition table	Including items such as:
	Inclusion and exclusion criteria in the sequence they were applied to the data
	Number of patients after application of each criterion
Describing patient characteristics of overall population	Including items such as:
	Number of patients
	N/% or mean (SD) of baseline characteristics
Describing outcomes and follow-up in overall population	Including items such as:
	Person-years of follow-up
	Mean, median follow-up time
	Reasons for censoring with numbers of subjects censored
	Number of health outcomes of interest
	Risk per 1000 persons
	Rate per 1000 person-years
B. Reporting of comparative results should include:	
Comparing patient characteristics for each exposure group	Including items such as:
	Number of patients
	N/% or mean (SD) of patient characteristics
	Absolute or standardized differences for compared groups
	Mahalanobis distance
Describing outcomes and follow-up for each exposure group	Including items such as:
	Person-years of follow-up
	Mean, median follow-up time
	Reasons for censoring with numbers of subjects censored
	Number of health outcomes of interest
	Risk per 1000 persons
	Rate per 1000 person-years
Relative measure of association (ratio)	Including items such as:
	Unadjusted and adjusted results
	Prespecified subgroup analyses
Absolute measure of association (difference)	Including items such as:
	Unadjusted and adjusted
	Pre-specified subgroup analyses
Additional diagnostic results when propensity score is used	Including items such as:
	Figure with propensity score distribution pre- and post-matching
	Tables for unmatched and matched population characteristics
	Tables for stratified population characteristics
	Tables for unweighted and weighted population characteristics
	Mean and distribution of weights
	N/% contributing to matched, trimmed, truncated or weighted analyses

(Continues)

TABLE 3 (Continued)

	Description
Additional diagnostic results when instrumental variable analysis is used	Table with distribution of population characteristics across levels of instrument
	Table with distribution of outcomes across levels of instruments
	Strength of association between instrument and exposure
	(eg, odds ratio, risk difference, partial R^2)
	Results of falsification tests:
	assumption that instrument does not affect outcome except through treatment
	assumption that instrument and outcome do not have common causes
C. Reporting of risk-adjustment methods should include:	
Estimand	What is being estimated with the risk-adjusted analytic method?
	(eg, ATT, ATE, marginal vs. conditional effect)
Measures of variability due to chance	How are standard errors obtained? (eg, model-based, bootstrap, robust)
Methods used for confounder adjustment:	
Direct or indirect standardization	What is the standard (reference) population?
	What covariates are used for standardization?
Stratification (on 1 or more covariates)	Which covariates define strata?
Multivariable outcome regression model	What kind of model was used? (eg, survival, binary, Poisson)
	Which covariates were used and how did they enter the model?
	(eg, binary, categorical)
Propensity score model	What kind of model was used? (eg, logistic, multinomial)
	Which covariates were used and how did they enter the model?
	(eg, binary, categorical)
If PS-matching	What matching algorithm, what caliper and on what scale?
	(eg, 0.025 standard deviations on the probability scale)
	What matching ratio? (eg, fixed 1:1, variable 1:5)
If PS-stratification	How are strata defined?
	(eg, deciles, centiles calculated among the exposed)
	Is trimming implemented before or after strata definition?
If PS-weighting	How are the weights calculated?
	Are the weights trimmed, truncated or stabilized?
Instrumental variable analysis	What kind of model was used (eg, two-stage least squares)
Matching	If the design involved matching, how did the analysis account for matching factors?

Abbreviations: ATE, average treatment effect; ATT, average effect among treated; PS, propensity score.

about how and whether the propensity score is trimmed or truncated prior to defining strata or weights. For stratification methods, the method for defining strata should be clearly defined; for example, are deciles of the propensity score constructed in the exposed group only or based on the distribution of the propensity score across the entire study population. For weighting methods, investigators should report how weights were calculated (numerator and denominator), whether the weights were trimmed, truncated or stabilized, and the mean and range of the weights.

When instrumental variable analysis is used for adjustment of confounding, investigators should report diagnostics from checking each of the three main assumptions for unbiased results. These include results from falsification tests of the assumption that the instrument affects outcome only through exposure, the strength of the relationship between the instrument and exposure, and potential confounders between the instrument and the outcome.[39,40]

For all RWE studies conducted using RWD, the statistical software programs, packages or platforms used in cohort extraction and

analysis should be reported (Table 2D). When relevant, the specific software version and macro or function settings or input parameters should be provided.

Items from Tables 1 and 2 have been incorporated into the revised RECORD-PE checklist.[41,42]

5 | REGISTRATION OF RWE STUDIES

Recent legislation (21st Century Cures, PDUFA VI, Adaptive Pathways) highlights the increasing focus on the potential use of RWE to support regulatory, reimbursement and other clinical decision-making. While there has been substantial successful experience with pharmacovigilance and post-approval safety studies using RWD[43,44] and recent growth in RWE studies of drug effectiveness,[45-47] there are different issues and implications for studies evaluating safety versus effectiveness. For example, there may be elevated concerns about financial and other incentives contributing to reporting of cherry-picked results when making secondary use of existing RWD in support of new indications. 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. Several options are available for registration of observational studies, including the EU Post-authorization Study Register, hosted by the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP), and two registries created by the National Institutes of Health's National Library of Medicine, ie, ClinicalTrials.Gov and HSRProj.[21]

6 | PRACTICAL ISSUES

Journals have word limits that make it difficult to provide important information about the complex decisions involved in RWE studies to minimize bias when making secondary use of healthcare data that were not collected for research purposes. Almost all journals allow online appendices and, alternatively, supplemental materials can be published on pre-publication websites such as medrxiv.org that can be referenced. We suggest that a high-level summary of methodology provided in manuscript text should be accompanied by key details of RWE study conduct, either as appendices or via citation of a registered protocol.

Transparency requires clarity of communication between researcher and reviewer. Providing reams of incomprehensible materials would not be transparent. Following the structure of the catalogue of parameters outlined in section III above could facilitate clarity in communication of key details. Use of a design diagram and provision of a table of design and analysis parameters to communicate methodology could reduce misinterpretation as well as reduce the number of words used in the methods section of the manuscript. Unambiguous reporting on RWE study conduct may in the future involve more standardized reporting formats, similar to randomized clinical trials, so that reviewers will know where to find the information they are looking for and more easily evaluate validity and compare across studies.

7 | DISCUSSION

Transparency of RWE studies is critical to understanding how findings of a specific study were derived and is a necessary foundation to assessing validity and determination of whether decisions should be informed by the findings.

In the present paper, we have reminded readers of the utility of thinking about emulation of a target trial when planning and communicating about RWE study implementation. We have recommended use of a graphical representation showcasing temporality of key longitudinal study design choices. We have highlighted study elements that should be reported to provide the clarity necessary to make a study reproducible. Finally, we have suggested registering study protocols to increase process transparency. With these tools, should they be used, the readership of RWE studies will be able to more efficiently understand each study and more able to assess a study's validity with reasonably high confidence before making decisions based on its findings.

Real-world evidence in diabetes research has many examples of misleading findings that are frequently flawed by study design biases, including immortal-time bias and adjustment for causal intermediates.[16,48-51] There are also recent prominent examples of studies showing unrealistically large survival benefits that can be largely explained by design biases.[1,52] Improved transparency may have brought many of these issues to the attention of journal editors and reviewers during the editorial review process. Most importantly, conceptualizing a target trial before conducting the research would probably have avoided these biases.

The RWE community has recognized lack of transparency as a key barrier to building confidence in RWE study findings and is working with regulators and journal editors to improve the transparency and interpretability of RWE studies. There is clearly room for improvement also in the context of diabetes RWE studies. Diabetes researchers can join the efforts of the wider RWE community and use the tools outlined in this paper to improve transparency in the conduct and reporting of RWE research in diabetes.

CONFLICT OF INTEREST

E.P. is investigator of investigator-initiated grants to the Brigham and Women's Hospital from GSK and Boehringer Ingelheim, not directly related to the topic of the submitted work. S.S. is an investigator of investigator-initiated grants to the Brigham and Women's Hospital from Bayer, Vertex, and Boehringer Ingelheim unrelated to the topic of this study. He is a consultant to WHISCON and to Aetion, a software manufacturer of which he owns equity. S.W.V. is investigator of investigator-initiated grants to the Brigham and Women's Hospital from Boehringer Ingelheim, Novartis Pharmaceuticals and Johnson & Johnson. These interests were declared, reviewed, and approved by the Brigham and Women's Hospital and Partners HealthCare System in accordance with their institutional compliance policies.

AUTHOR CONTRIBUTIONS

Conception and design: E.P., S.S., S.V.W. Analysis and interpretation of the data: E.P., S.S., S.V.W. Drafting of the article: E.P., S.S., S.V.W. Critical revision of the article for important intellectual content: E.P., S.S., S.V.W. Final approval of the article: E.P., S.S., S.V.W.

ORCID

Elisabetta Patorno https://orcid.org/0000-0002-8809-9898

REFERENCES

1. Suissa S. Lower Risk of Death With SGLT2 Inhibitors in Observational Studies: Real or Bias? *Diabetes Care*. 2018;41(1):6-10.
2. Patorno E, Goldfine AB, Schneeweiss S, et al. Cardiovascular outcomes associated with canagliflozin versus other non-gliflozin antidiabetic drugs: population based cohort study. *BMJ*. 2018;360:k119.
3. Pasternak B, Ueda P, Eliasson B, et al. 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. Grodstein F, Manson JE, Colditz GA, Willett WC, Speizer FE, Stampfer MJ. A Prospective, Observational Study of Postmenopausal Hormone Therapy and Primary Prevention of Cardiovascular Disease. *Ann Intern Med*. 2000;133(12):933-941.
5. 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.
6. Chan KA, Andrade SE, Boles M, et al. Inhibitors of hydroxymethylglutaryl-coenzyme A reductase and risk of fracture among older women. *Lancet*. 2000;355(9222):2185-2188.
7. 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.
8. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? *Clin Pharmacol Ther*. 2017;102(6):924-933.
9. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidimiology*. 2008;19(6):766-779.
10. Kosiborod M, Cavender MA, Fu AZ, et al. Lower Risk of Heart Failure and Death in Patients Initiated on SGLT-2 Inhibitors Versus Other Glucose-Lowering Drugs: The CVD-REAL Study. *Circulation*. 2017;136:249-259.
11. Udell JA, Yuan Z, Rush T, Sicignano NM, Galitz M, Rosenthal N. Cardiovascular Outcomes and Risks After Initiation of a Sodium Glucose Co-Transporter 2 Inhibitor: Results From the EASEL Population-Based Cohort Study. *Circulation*. 2018;137(14):1450-1459.
12. 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.
13. Wang SV, Schneeweiss S, Berger ML, et al. Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0. *Pharmacoepidemiol Drug Saf*. 2017;26(9):1018-1032.
14. Meier CR, Schlienger RG, Kraenzlin ME, Schlegel B, Jick H. HMG-CoA reductase inhibitors and the risk of fractures. *JAMA*. 2000;283(24):3205-3210.
15. Smeeth 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.
16. Suissa S, Azoulay L. Metformin and the risk of cancer: time-related biases in observational studies. *Diabetes Care*. 2012;35(12):2665-2673.
17. Schneeweiss S, Huybrechts KF, Gagne JJ. Interpreting the quality of health care database studies on the comparative effectiveness of oral anticoagulants in routine care. *J Comp Eff Res*. 2013;3(4):33-41.
18. de Vries F, de Vries C, Cooper C, Leufkens B, van Staa T-P. Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: the General Practice Research Database. *Int J Epidemiol*. 2006;35(5):1301-1308.
19. 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 Mar;99(3):325-332.
20. Framework for FDA's Real World Evidence Program. U.S. Food & Drug Administration. 2018. https://www.fda.gov/media/120060/download. Accessed November 25, 2019.
21. Berger ML, Sox H, Willke RJ, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiol Drug Saf*. 2017;26(9):1033-1039.
22. Schneeweiss S, Rassen JA, Brown JS, et al. Graphical Depiction of Longitudinal Study Designs in Health Care Databases. *Ann Intern Med*. 2019;170(6):398-406.
23. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol*. 2016;183(8):758-764.
24. Hernán 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.
25. Layton JB, Kshirsagar AV, Simpson RJ Jr, et al. Effect of statin use on acute kidney injury risk following coronary artery bypass grafting. *Am J Cardiol*. 2013;111(6):823-828.
26. Kim SC, Solomon DH, Rogers JR, et al. Cardiovascular Safety of Tocilizumab Versus Tumor Necrosis Factor Inhibitors in Patients With Rheumatoid Arthritis: A Multi-Database Cohort Study. *Arthritis Rheumatol*. 2017;69(6):1154-1164.
27. Bykov K, Schneeweiss S, Glynn RJ, Mittleman MA, Bates DW, Gagne JJ. Updating the Evidence of the Interaction Between Clopidogrel and CYP2C19-Inhibiting Selective Serotonin Reuptake Inhibitors: A Cohort Study and Meta-Analysis. *Drug Saf*. 2017;40(10):923-932.
28. Brookhart MA. Counterpoint: the treatment decision design. *Am J Epidemiol*. 2015;182(10):840-845.
29. Douglas IJ, Langham J, Bhaskaran K, Brauer R, Smeeth L. Orlistat and the risk of acute liver injury: self controlled case series study in UK Clinical Practice Research Datalink. *BMJ*. 2013;346:f1936.
30. Toh S, Reichman ME, Houstoun M, et al. Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. *Arch Intern Med*. 2012;172(20):1582-1589.
31. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care*. 2013;51(8 Suppl 3):S22-S29.
32. Kahn MG, Brown JS, Chun AT, et al. Transparent reporting of data quality in distributed data networks. *EGEMS*. 2015;3(1):1052.
33. Langan SM, Benchimol EI, Guttmann A, et al. Setting the RECORD straight: developing a guideline for the REporting of studies Conducted using Observational Routinely collected Data. *Clin Epidemiol*. 2013;5:29-31.
34. Ray WA. Evaluating Medication Effects Outside of Clinical Trials: New-User Designs. *Am J Epidemiol*. 2003;158(9):915-920.
35. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf*. 2010;19(6):537-554.

36. 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.
37. Naing NN. Easy Way to Learn Standardization : Direct and Indirect Methods. *Malays J Med Sci*. 2000;7:10-15.
38. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*. 2011;46(3):399-424.
39. Swanson SA, Hernán MA. Commentary: How to Report Instrumental Variable Analyses (Suggestions Welcome). *Epidemiology*. 2013;24(3):370-374.
40. Glymour MM, Tchetgen Tchetgen EJ, Robins JM. Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions. *Am J Epidemiol*. 2012;175(4):332-339.
41. Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ*. 2018;363:k3532.
42. Langan SM, Schmidt SAJ, Wing K, et al. La déclaration RECORD-PE (Reporting of Studies Conducted Using Observational Routinely Collected Health Data Statement for Pharmacoepdemiology) : directives pour la communication des études realisées à partir de données de santé observationelles collectées en routine en pharmacoépidémiologie. *Can Med Assoc J*. 2019;191(25):E689-E708.
43. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative–A comprehensive approach to medical product surveillance. *Clin Pharm Ther*. 2016;99(3):265-268.
44. Suissa S, Henry D, Caetano P, et al. CNODES: the Canadian Network for Observational Drug Effect Studies. *Open Med*. 2012;6(4):e134-e140.
45. Toh S, Hampp C, Reichman ME, et al. Risk for Hospitalized Heart Failure Among New Users of Saxagliptin, Sitagliptin, and Other Antihyperglycemic Drugs: A Retrospective Cohort Study. *Ann Intern Med*. 2016;164(11):705-714.
46. Azoulay L, Filion KB, Platt RW, et al. Association Between Incretin-Based Drugs and the Risk of Acute Pancreatitis Incretin-Based Drugs and the Risk of Acute PancreatitisIncretin-Based Drugs and the Risk of Acute Pancreatitis. *JAMA Intern Med*. 2016;176(10):1464-1473.
47. Filion KB, Azoulay L, Platt RW, et al. A Multicenter Observational Study of Incretin-based Drugs and Heart Failure. *N Engl J Med*. 2016;374(12):1145-1154.
48. Patorno E, Patrick AR, Garry EM, et al. Observational studies of the association between glucose-lowering medications and cardiovascular outcomes: addressing methodological limitations. *Diabetologia*. 2014;57(11):2237-2250.
49. Patorno E, Garry EM, Patrick AR, et al. Addressing limitations in observational studies of the association between glucose-lowering medications and all-cause mortality: a review. *Drug Saf*. 2015;38(3):295-310.
50. 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.
51. Garry EM, Buse JB, Gokhale M, et al. 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.
52. Suissa S. Reduced Mortality With Sodium-Glucose Cotransporter-2 Inhibitors in Observational Studies: Avoiding Immortal Time Bias. *Circulation*. 2018;137(14):1432-1434.
53. 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;42:2204-2210.

How to cite this article: 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):45–59. https://doi.org/10.1111/dom.13918