Building reliable evidence from real-world data: methods, cautiousness and recommendations

Giovanni Corrao[1]

ABSTRACT

Routinely stored information on healthcare utilisation in everyday clinical practice has proliferated over the past several decades. There is, however, some reluctance on the part of many health professionals to use observational data to support healthcare decisions, especially when data are derived from large databases. Challenges in conducting observational studies based on electronic databases include concern about the adequacy of study design and methods to minimise the effect of both misclassifications (in the absence of direct assessments of exposure and outcome validity) and confounding (in the absence of randomisation). This paper points out issues that may compromise the validity of such studies, and approaches to managing analytic challenges. First, strategies of sampling within a large cohort, as an alternative to analysing the full cohort, will be presented. Second, methods for controlling outcome and exposure misclassifications will be described. Third, several techniques that take into account both measured and unmeasured confounders will also be presented. Fourth, some considerations regarding random uncertainty in the framework of observational studies using healthcare utilisation data will be discussed. Finally, some recommendations for good research practice are listed in this paper. The aim is to provide researchers with a methodological framework, while commenting on the value of new techniques for more advanced users.

Key words: Databases; Medical records; Observational studies; Pharmacoepidemiology; Record linkage

(1) Department of Statistics and Quantitative Methods, Division of Biostatistics, Epidemiology and Public Health, University of Milano-Bicocca, Milan, Italy

CORRESPONDING AUTHOR: *Giovanni Corrao, Department of Statistics and Quantitative Methods, Division of Biostatistics, Epidemiology and Public Health, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, Building U7, 20126 Milan, Italy. Tel: +39 02 64485854. Fax: +39 02 64485899. e-mail: giovanni.corrao@unimib.it*

DOI: ***10.2427/8981***

1. PRELIMINARY REMARKS

In the last decades, routinely collected and electronically stored healthcare information has spread throughout the world [1]. Large computerised databases with millions of observations regarding the use of drugs, vaccines, devices, and procedures along with health outcomes may be useful in assessing which treatments are most effective and safe in routine care without long delays and the prohibitive costs of randomised clinical trials (RCTs) [2]. There is, however, some reluctance on the part of many health professionals to use observational data, especially when data are derived from large databases [2]. This distrust issues, at least in part,

from the fact that observational studies have been downgraded in the hierarchy of evidence [3], primarily because of their heterogeneity, the potential for systematic and random errors, as well as discordance between some observational studies and RCTs [4, 5]. Challenges in conducting observational studies based on electronic databases encompass concern about the adequacy of study design, approaches to minimise the effect of misclassification in the absence of direct assessments of exposure and outcome validity, and control of confounding in the absence of randomisation [6-12]. Such threats to validity limit the usefulness of these studies and application of findings in policy and practice. However, with proper research design and application of an array of traditional and newer analytical approaches such matters of concern may be addressed to improve our understanding of care effects [13, 14].

This paper outlines issues that may compromise the validity of such studies, and approaches to manage such analytical challenges. First, a strategy of sampling within a large cohort, as an alternative to analysing the full cohort, will be presented (par. 2). This sampling scheme is crucial for observational studies that use real-world data, where cohorts are large and formidable challenges in data analysis are consequently required [7]. Second, methods for controlling outcome and exposure misclassifications will be described (par. 3). Methods to overcome measuring errors must be implemented particularly when, owing to the lack of relevant information, approximations of the level of exposure to medical care are necessarily used [15]. Third, several techniques that take into account both measured and unmeasured confounders will also be presented (par. 4). These techniques play a crucial role when data is collected from real-world databases, especially healthcare utilisation (HCU) data, where detailed clinical, lifestyle, and socioeconomic information is systematically lacking [16]. Fourth, some considerations regarding random uncertainty in the framework of observational studies using HCU data will be discussed (par. 5). Finally, some recommendations for good research practice in this setting are listed in paragraph 6.

The goal is to provide researchers with a methodological framework and to supply a critical overview of the new techniques for more advanced users.

2. DESIGNING A STUDY

2.1. Basic cohort design

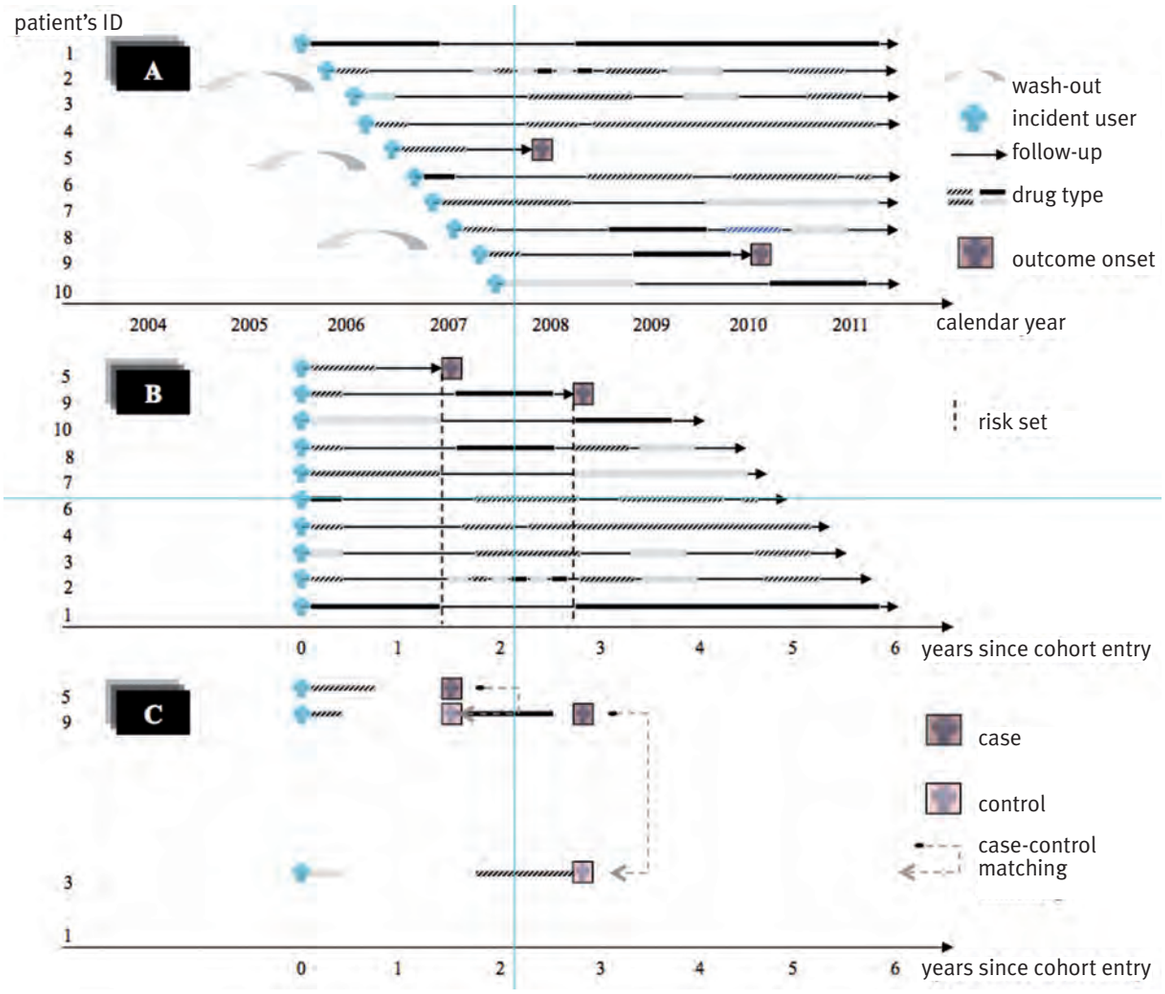
As an illustrative example, we describe a cohort of ten incident users (inception cohort) of oral hypoglycaemic drugs, ranked according to the date of first prescription of the considered drugs (according with Ray [17], this approach has been described “new user design” already in the introductory article of this issue). Figure 1, box A, clarifies that this cohort can be obtained by considering the first drug dispensation during years 2006-2007 and by excluding patients with at least one prior dispensing oral hypoglycaemic drug. This approach is comparable with the wash-out period approach commonly used in RCTs [13]. Other restriction criteria for cohort entry can be used but must be clearly reported in the study protocol. For example, younger patients and those who previously experienced the outcome of interest might be excluded in order to study patients more likely affected by type 2 diabetes and to focus on primary prevention of the considered outcome.

Figure 1, box A, also depicts the heterogeneity of exposure for both drug type and treatment duration. This is *per se* an important finding in the setting of real-world clinical practice. Evaluating deviations from guideline-based medical recommendations, assessing persistence with, and adherence to oral hypoglycaemic drugs, and measuring costs linked with diabetes treatment, may be easily achieved by means of this new-user design.

Arguably, the most important and challenging part in using this basic cohort design is to detect causes of heterogeneity that can affect the outcome. To better illustrate this new perspective, an alternative depiction of the same cohort is reported in Figure 1, box B, which presents a new time axis considering the years since the start of treatment, ranking subjects according to the length of follow-up time. Hence, it is clear that patients who experienced the outcome had a shorter time of drug exposure than those who did not experience it. Indeed, every patient belonging to the first category was treated in average only one year against more than three years benefitted by outcome-free patients. On the other hand, also considering the shorter follow-up period of patients who experienced the

FIGURE 1

ILLUSTRATION OF A FIXED COHORT OF TEN INCIDENT USERS OF A GIVEN THERAPY GENERATED FROM A WELL DEFINED DYNAMIC POPULATION FROM 2006 THROUGH 2007



Cohort members are monitored until 31 December 2011. Calendar time and time since cohort entry are used as time axes, as illustrated in box A and box B, respectively. Case-control sample of one control per case nested into the same cohort is illustrated in box C

outcome, this group resulted from drug therapy for about 50% of the follow-up time, against 61% of those who benefitted by the therapy but did not experience it. Accordingly, one might conclude that drug exposure prevents onset of the outcome. This is, however, a misleading conclusion because it was drawn regardless of the time varying nature of drug treatment.

A tool that may help us to illustrate the follow-up cohort experience is the risk set. A risk set is formed by all members of the cohort who are at risk of the outcome at a given point in time, that is, they are free of the outcome of interest and are members of the cohort at that point in time. The only relevant risk sets for data analysis are those defined by the time of occurrence of each outcome [7]. The reader can verify that before the onset of the first outcome (first risk set), drug therapy was available for about 8 months (covering nearly 48% of the time of follow-up) for both the single patient who experienced the outcome (i.e. patient # 5), as well as the average follow-up time of the remaining nine outcome-free patients. As far as the second risk set is concerned, drug therapy was available for about 16 months (and covered nearly 53% of the follow-up time) for both the single patient who experienced the

second outcome (i.e. patient # 9), as well as, in average, the remaining eight outcome-free patients. Hence, the time of occurrence of an outcome does not seem to be associated with drug exposure.

The example suggests that, when studying exposures that vary with time, such as drug availability or healthcare room attendance, time-dependent exposure must be accounted for in both design and analysis [18, 19]. This can be accomplished by including time-dependent covariates in a Cox proportional-hazards regression model [20]. Briefly, with this approach, the hazard ratio associated with exposure is derived by using information that is concurrent to the observed outcomes, rather than the exposure profile over the full length of the follow-up period [21]. This approach, however, requires enormous computational resources especially when real-world HCU data are used, and very large cohorts are accordingly formed.

2.2. Sampling within the cohort

Alternatively, the nested case-control approach may be used provided that the exposure information among controls (patients who do not experience the outcome) reflects values corresponding to the time of selection of their respective case (patients who experience the outcome). This may be accomplished by including in the study, along with all patients who experience an outcome (cases), also a random sample of those who did not experience it (controls).

The idea of a nested case-control design within a cohort was first introduced in 1973 by Mantel [22], who proposed an unmatched selection of controls and called it a synthetic retrospective study. The nested case-control design involves three steps: (i) selecting all cases in the cohort; (ii) forming the risk set corresponding to each case; and, finally, (iii) randomly selecting one or more controls from each risk set [7].

Figure 1, box C, illustrates the selection of a case-control sample nested into the above-described cohort, with one control per case (1:1 matching). It is clear from the definition of risk set, that a future case is eligible to be a control for a prior case, as illustrated in Figure 1, box C, for patient # 9, and that a subject may be selected as a control more than once. If, instead, controls are necessarily selected only from non-cases and subjects are not permitted to be considered more than once in the nested case-control sample, a bias is introduced in the estimation of the relative risk because control exposure prevalence will lean towards subjects with a longer follow-up period who do not become cases during the study follow-up [7].

The issue of required criteria for case-control matching is important. It is clear that members of each case-control(s) set must have the same observational time-window length; that is, all members would have experienced the same exposure pattern under the null hypothesis (i.e., absence of exposure-outcome association). This was achieved, according to the concept of risk set, by randomly selecting as controls patients who were still at risk of developing the outcome at the time when the index case had the event, as illustrated in Figure 1, box C. In pharmacoepidemiology and healthcare research, however, drug exposure can vary substantially over calendar time, thus introducing a "cohort effect". In order to avoid it, controls would be matched at the index case also by start date of treatment (follow-up). Other matching criteria, although important for some research issues, are not so essential since matching criteria are an alternative to adjustments made by data analysis for the corresponding effects (see par. 4.3.2).

Another important issue is the required number of controls per case. Since the number of cases in the cohort is fixed and cannot be increased to satisfy this requirement, the only remaining alternative is to increase the control-to-case ratio. It can be readily noticed in sample size tables that the power significantly rises with every additional control up to four controls per case, but becomes negligible beyond this ratio [the reader can easily see the following website for details of the sample size formula: http://www.statsdirect.co.uk/help/sample_size/ssmc.htm]. Although this general rule of an optimal 4:1 control-to-case ratio is appropriate in the majority of cases, one should be aware that when drug exposure is infrequent, the theorised relative risk moves beyond unity, or several factors or other drugs are being assessed simultaneously, the ratio could easily be required to increase to 10 or more controls per case [7]. This is generally not a problem in HCU-based investigations.

The analysis of data from a nested case-control study must maintain the time-matched nature of the selection of cases and controls. Conditional logistic regression is the appropriate model as it uses case-controls set as the fundamental unit of analysis, in agreement with the proportional hazards model of the full cohort.

It has been recently proven empirically that a nested case-control approach can be used to analyse a cohort with time-dependent covariates, with results that are similar to those obtained by Cox regression [23]. Additionally, given that the nested case-control approach obviates the computationally intensive calculations involved in Cox regression when time-dependent covariates are used, the paper also illustrates the large quantitative reduction in CPU time required for analysis. This explains why the nested case-control design is an increasingly popular sample technique to address issues of time-dependent exposures, which are commonly encountered in HCU-based studies that estimate drug and other healthcare effects.

2.3. Other observational designs

Another well recognised observational design implying sampling within a cohort is the one we currently call case-cohort design. It was originally used by Hutchison [24] in performing external comparisons of leukaemia rates in patients submitted to radiation therapy for cervical cancer. It was ultimately developed and formalised by Prentice [25], who coined the name "case-cohort". In the case-cohort design, the cohort is defined as in a cohort study (par. 2.1), in which cohort members attend a follow-up period to identify those who experience the outcome. If additional information is required (e.g., blood level of a given disease marker, or biological material for genetic analyses), then said information may be collected for all cases, and for a random sample selected from the included cohort. Note that, conversely to the nested case-control design, which considers the exposure experience from entry until onset of the outcome, the case-cohort design is intended to supply additional information characterising the cohort members when they joined the cohort. Compared to the cohort design, the case-cohort one is intended to increase efficiency when additional information needs to be collected or when studying more than one outcome [26].

Some additional designs like 2-stage studies [27], validation studies [16], and case-only designs [28] are not discussed here, since they have been created to deal with issues regarding confounding under specific scenarios. Some of these issues will be discussed in the following paragraphs as they are useful to deal with confounding.

3. MEASUREMENT ERRORS AND MISCLASSIFICATION

Errors can occur in measuring both exposure and outcome. These errors lead to classification bias, that is (i) identifying subjects as having experienced the disease outcome when they have not (or being exposed to a drug when they are not), or (ii) not having experienced the outcome when they experienced it (or not being exposed when they are).

Classification bias is further categorised as differential or non-differential. Differential misclassification is present when the likelihood of outcome misclassification differs between exposed and unexposed (or exposure misclassification differs between subjects who experience outcome and those who do not). An example of differential exposure misclassification is when exposed patients have a lower likelihood of outcome misclassification because, since they have to enter the healthcare system to receive medication, their likelihood of recording a correct diagnosis increases. Those not exposed are much more likely to be misclassified as not having the disease, which is an artefact of not entering the healthcare system [13]. Non-differential misclassification occurs when the likelihood of misclassification is the same across either the exposed or outcome groups. Examples of non-differential misclassifications include those caused by coding errors of diagnosis or medical procedure reported in the hospital discharge database (e.g., according to ICD 9th revision), medicaments reported in drug prescription database (e.g., in the ATC code), or identifier code (e.g., in the fiscal or health individual code) due to accidental mistyping. Upcoding, assigning codes of higher reimbursement value over codes with lesser reimbursement value, is

an additional source of error at coder level [29].

The direction of misclassification regarding measures of association will depend on the type of misclassification [30]; when it is differential, association measures would be biased either toward or away from the null [31]. The effect of non-differential misclassification varies by the factor that is misclassified. When outcome variables suffer of non-differential misclassification, association measures are typically biased toward the null [32]. When non-differential misclassification affects variables used to define the cohort (e.g., by excluding prevalent disease at baseline), association measures could be biased away from the null, with the degree of bias varying by disease incidence and prevalence [33]. Therefore, the effect of misclassification on study conclusions even when a single variable alone is misclassified is unpredictable.

3.1. Outcome misclassification

Specific disease coding must be used to extract patients experiencing a given outcome and to characterise their clinical profile from HCU data.

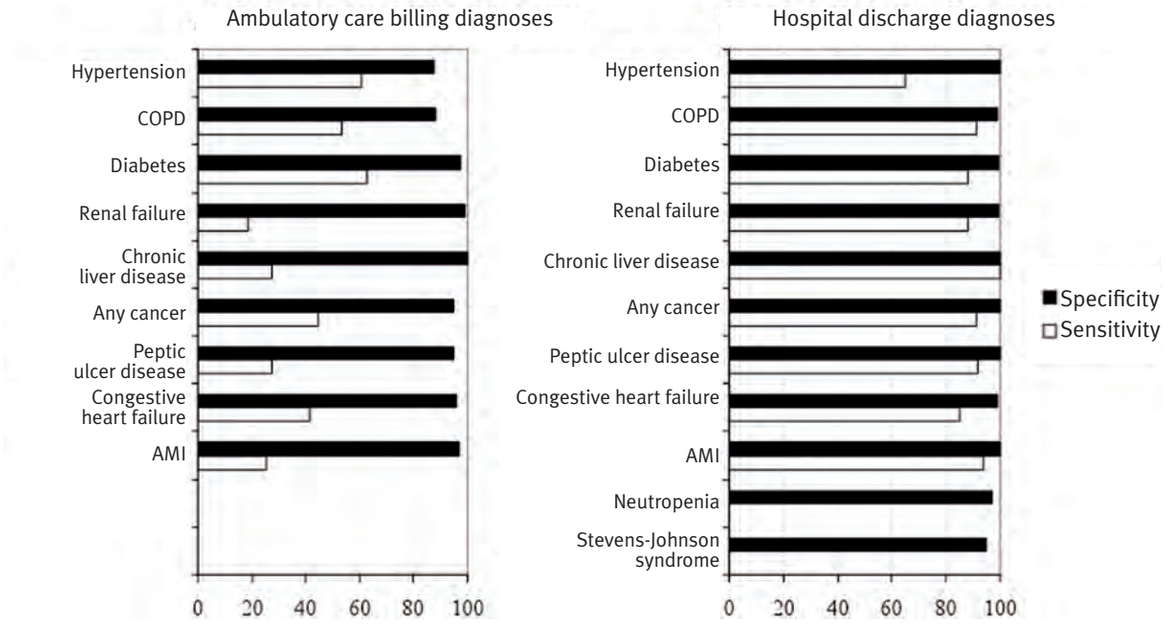
The International Classification of Diseases, 9th revision, Clinical Modification: ICD-9-CM (http://icd9cm.chrisendres.com/), and the ICD-10th revision in some countries (http://apps.who.int/classifications/icd10/browse/2010/en), are the classification systems of disease and medical procedures, common to most HCU databases. Each disease coding system necessarily implies a loss of information as a simple consequence of classification, as opposed to nomenclature. Many authors have discussed the qualitative loss that occurs with ICD-9-CM coding [34-39]. Besides problems with billing considerations distorting coding, and the errors of coding caused thereby, they note that coding diagnosis categories lose essential information about the true conditions of patients [40].

To understand the effect of outcome misclassification on association measure estimates, it is important to note that a lack of specificity of the outcome measurement is worse than a lack of sensitivity in most situations. If specificity of the outcome assessment is 100%, then relative risk estimates are unbiased [41]. Given this, literature on misclassification of HCU data diagnoses is not as depressing as it might first appear to be. A recent comprehensive study on the misclassification of HCU data diagnoses using medical records review as the gold standard revealed that the sensitivity of claims diagnoses is often less than moderate, whereas specificity is usually 95% or greater (Figure 2) [42]. A high specificity of diagnostic coding in HCU data can be expected because if a diagnosis is coded and recorded, this diagnosis was most likely made, particularly in hospital discharge summaries [43]. From this perspective, HCU data may be assumed to be suitable for association studies (e.g., to investigate the effect of healthcare on the onset of a given outcome), rather than to measure the incidence of diseases (e.g., incidence and prevalence).

Since chronic diseases usually require multiple contacts with the health system, a single diagnostic code may not suffice to accurately identify cases [44]. This explains the widespread use of diagnostic algorithms in identifying patients who experience a given outcome [45]. Furthermore, the use of hospital charts to identify cases limits the possibility of detecting all the outcomes (e.g., those that do not require hospital admission), thus introducing a bias due to the selective inclusion of more severe outcomes. Alternative techniques have been specifically developed for detecting less severe outcomes. For example, when a drug A is suspected of causing an outcome that itself is treated by a drug B, we only need to search the drug prescription database for patients who experience outcome. The effect of antibiotics on the risk of arrhythmias has been investigated with this method in a study that estimated the association between use of drugs belonging to the classes of antibacterials (exposure) and the risk of arrhythmia (outcome) [46]. Yet, the suspected causal association between the use of statins and the risk of diabetes [47] might be investigated by studying the strength of the association between statin use (exposure) and the risk of starting an antidiabetic therapy (outcome). It should be considered, however, that patients on statins are those in whom diabetes is most likely diagnosed. In other words, we cannot exclude the possible occurrence of a detection bias [48] with a subsequent amplification in the exposure-outcome relationship.

FIGURE 2

SENSITIVITY AND SPECIFICITY OF SELECTED DIAGNOSES IN DATABASES FROM HOSPITAL AND CLINICAL CARE



AMI, acute myocardial infarction; COPD, chronic obstructive pulmonary disease

Source: Schneeweiss S & Avorn J [6]

3.2. Exposure misclassification

The drug coding system, i.e. the Anatomical Therapeutic Chemical (ATC) classification system of the WHO (http://www.whocc.no/atc_ddd_index/), must be used for extracting patients who use a given medicament, as well as those who use other drugs and can be considered as proxies of their clinical profile, from HCU data.

Pharmacists fill out drug prescriptions with little room for interpretation and are reimbursed by Health Authorities on the basis of detailed, complete, and accurate claims that are submitted electronically [49-51]. Pharmacy dispensing information is, therefore, expected to provide highly accurate data, also because filling out an incorrect report about dispensed drugs has legal consequences [52]. These data are usually seen as the gold standard of drug exposure information compared to self-reported information [53] or prescribing data in outpatient medical records [54].

The time-window during which patients are considered exposed to a given drug is often a key variable in assessing exposure [13]. This variable requires at least that information for calculating the days covered by each prescribed drug canister are either available, or may be approximated. Within most USA prescription databases, the "days of supply" are usually included as a data field within each prescription claim (e.g., 60 tablets of a medication that is taken twice daily would yield a 30-day supply), along with the dispensation dates of the prescription [55]. In several non-USA databases (e.g., in Italy), information about date dispensation and dispensed quantity are available, but not those about the prescribed drug dose. Hence the need to approximate the number of days covered by each dispensed canister by considering the quantity of drug contained in the dispensed canister and some metric calculations of the average daily dose, e.g., defined daily dose (DDD), established as the typical adult's daily maintenance dose [56]. However, it should be emphasised that the DDD is a unit of measurement, and does not necessarily mirror the recommended or Prescribed Daily Dose (PDD). Discrepancies between PDD (which may by assumed as a gold standard of exposure duration), and DDD

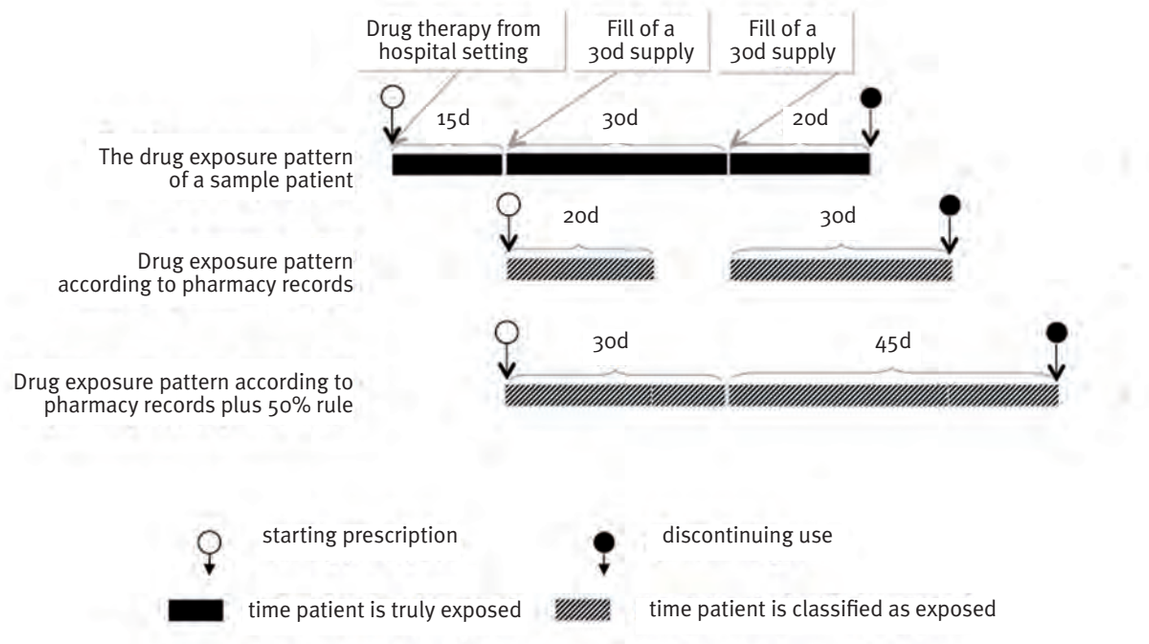
(which is likely affected by measurement error) necessarily introduce classification bias.

Figure 3 illustrates some typical misclassification problems in drug exposure assessment [6]. If the calculated days of supply are too few, or if patients decide to stretch a prescription by using a lower dose (e.g., tablet splitting), some person-time will be classified as unexposed when actually exposed. Most chronically administered drugs are used for longer periods, resulting in multiple refills. A patient can thus be classified as being unexposed intermittently despite continuous exposure. As a result, many investigators extend the calculated days of supply by a fraction (e.g., 50%) to avoid this misclassification. However, said strategy aggravates another misclassification that can occur if a patient discontinues drug use without finishing the supply. The right balance between improved sensitivity of drug exposure assessment and specificity depends on how precisely the days of supply are calculated; this depends on the type of drug and how regularly it is taken.

Over-the-counter (OTC) medications present a scenario in which misclassification is particularly problematic [57]. Measurements based on HCU data will exclude the use of OTC products and lead to misclassification of exposure to those medications [58, 59]. The inability to measure exposure during the observation period can also be an issue if the available data do not fully capture all sources of exposure. The use of OTC medication as exposure is one example of not being able to accurately capture all exposures, but this may occur in other circumstances. For example, in most databases, hospital discharge files do not contain drug use information. Therefore, every stay in hospital represents a period with missing drug exposure information by design, thus generating a particular form of exposure misclassification labelled as "immeasurable time bias" [60]. Furthermore, exposure may be misclassified because of protopathic bias, i.e. the drug use could be attenuated or interrupted owing to the onset of early symptoms of the outcome [61]. It must be said that both sources of bias likely generate differential exposure misclassification because they mainly affect patients who experience the outcome.

FIGURE 3

TYPICAL CAUSES FOR DRUG EXPOSURE MISCLASSIFICATION IN STUDIES BASED ON DRUG PRESCRIPTION DATABASES



Source: Schneeweiss S & Avorn J [6], modified

3.3. Strategies to account for misclassification

3.3.1. Algebraic methods

The impact of misclassification can be considerable but is rarely quantified. If the sensitivity and specificity of an exposure or outcome measurement are known, simple algebraic methods can be applied for unmatched [32] and matched analyses [62]. All these methods are, however, applied only to unadjusted associations (i.e., raw 2×2 data tables), which is an unrealistic analysis in pharmacoepidemiology and healthcare research. Other techniques use the positive predictive value (PPV) of the outcome measure to be determined in a separate validation study [63, 64]. This is an interesting approach for database studies because typically PPVs are easier to estimate in internal validation studies than sensitivity and specificity. This approach requires (i) a sample of records to be randomly selected from the HCU database; (ii) the corresponding medical documentation to be traced to the provider (i.e., the hospital) where the records were generated; (iii) medical documentation to be reviewed by a trained physician, possibly blinded regarding the diagnostic code as reported in the HCU database. However, because data anonymization is often employed by institutional review boards for protecting patient privacy, this approach might be unfeasible.

3.3.2. Sensitivity analyses

Because the lack of detailed information makes the use of approximations a necessity to define both outcome and exposure, the effect of such approximations can be explored by sensitivity analyses. Sensitivity analysis tests various criteria to define a critical measure in an effort to determine how those definitions affect the results and interpretation [65]. For instance, the influence of diagnostic criteria for outcome definition (e.g., using either ICD-9 codes or a combination of these codes and free-text, if available), or the length of the time-window for exposure definition (e.g., extending the calculation of day of supply adding different fractions to the calculation based on DDD) are commonly used techniques for investigating the robustness of findings by varying criteria applied in the principal analyses. Yet, the application of various methods for correcting immeasurable time bias [60], protopathic bias [61], and detection bias [67] would be used when appropriate.

Finally, a sensitivity approach for estimating the effect of OTC drugs on the exposure-outcome association has been recently proposed [68].

3.3.3. External adjustment

Methods that use individual validation data to account for measurement errors have been available for several years [69]. Among these, regression calibration (RC) may play an interesting role. RC is an intuitive, non-iterative statistical method that is useful for adjusting point and interval estimates obtained from regression models for measurement-error bias [70].

To illustrate the RC method, consider an HCU-based study measuring the effect of X' (e.g., the coverage length for a given drug dispensation approximated by DDD) and a dichotomous outcome Y (Y=1 for event, 0 for no event). The effect of X' on the risk of Y is usually obtained from the logistic-regression model

$$\text{logit}[P(Y=1 \mid X')] = \beta_0 + \beta' X'$$

When a gold standard assessment is available, a validation study conducted in a random subsample of patients included in the main (HCU) study can be used to validate the usual exposure measure (X') against its gold standard (X). For example, data obtained by a sample of physicians operating in the same area as the target population under study may be informative about the drug doses prescribed to patients. Assuming the prescribed daily dose as gold standard (X), and considering that X' is also measurable from this validating data source, the misclassification function may be easily obtained by the linear-regression model

$$E(X \mid X') = \alpha + \gamma X'$$

Rosner et al [71] proposed to estimate logarithm relative risk by substituting β' with

$$\hat{\beta} = \hat{\beta}' / \hat{\gamma}$$

the corresponding variance being

$$\text{var}(\hat{\beta}) = \frac{1}{\hat{\gamma}^2}\text{var}(\hat{\beta}') + \frac{\hat{\beta}'^2}{\hat{\gamma}^4}\text{var}(\hat{\gamma})$$

The RC method is, therefore, an appropriate

approach when information about true exposure is available from a validation study and the assumption of a linear relationship between observed and true exposure measures is not.

RC was originally proposed by Willett in a study aimed to measure the effects of fat and fibre on breast cancer [72], and has been widely used in the field of nutritional epidemiology [15, 73] as well as in other fields [74, 75]. Strangely, though RC does not involve particular computational complexities, it is not very popular in the field of pharmacoepidemiology and healthcare research.

3.4. Misclassification due to record linkage errors

Each HCU database supplies information about a single care setting (e.g., pharmacy and hospital files). Research methods, however, need to build a patient history that captures clinical encounters across several care settings. The development of methods to link records from different sources across time and providers has achieved this [76-78]. Record-linkage (RL) helps the identification of the same patient in various files containing different types of information. For example, by linking drug prescription, hospital discharge and vital statistics information we are able to generate a longitudinal electronic patient history, which allows assessing safety, effectiveness and cost of healthcare, and other health-related objectives [79].

Although RL is an important tool in observational research, it may be associated with various types of error. When linking two databases, there is a proportion of records that will match and a proportion that will remain unmatched. An error arises if data sources do not consistently capture the same cases, records that correspond to the same person fail to link due to missing or inaccurate data (false negatives), or unrelated records are mistakenly linked (false positives) [80].

The success of linkage, often described in terms of minimising mismatches, can depend on a number of factors, including the quality of the information used in the linkage process and how uniquely identified reported information is. Recent studies have shown that, unlike deterministic methods, the flexibility of probabilistic record linkage allows for minimisation of mismatches under variations in data quality [81]. A systematic review of the accuracy of probabilistic linkage identified six papers that had complete data on summary measures of linkage quality and found that linkage sensitivity (i.e., the proportion of truly matched records detected) ranged from 74% to 98%, and that specificity (i.e., the proportion of truly unmatched records detected) ranged from 99 to 100% [82].

At a glance, these arguments may be interpreted as a reason to implement probabilistic RL procedures when conducting observational studies based on HCU data. Caution is, however, recommended. First, probabilistic RL procedures require enormous computational resources, especially when real-world HCU data are used. For example, probabilistic RL between drug prescriptions (DP) and hospital admissions (HA) during a time window of five years among residents in the Italian Region of Lombardy (i.e., between the almost $387 \cdot 10^6$ DP, and $5 \cdot 10^5$ HA records over five years) requires the comparison of (DP • HA =) $193.5 \cdot 10^{12}$ pairs of identifier codes; each of them should be submitted to a decisional rule to either match or not match the corresponding pair of records! As a consequence, even admitting that sufficient computational resources are available, probabilistic RL procedures would be implemented when deterministic RL introduces important misclassification errors. It has been recently reported that, assuming true positive rates between 60% and 75%, the relative bias is about 5% for large study sizes as those based on HCU data [79].

Second, the rule for matching a couple of records according to probabilistic procedure is based on the discriminating power and accuracy of identifier codes (i.e., in what are the two codes alike?) [77]. Hence, similar, but not equal, identifiers (which should not be matched by a deterministic procedure) are attributed to the same patient if the probability that they belong to a single patient is higher than an established threshold (and the probability that they belong to different patients is lower than an established threshold), thus reducing the number of false negatives, and increasing the matching sensitivity. On the other hand, this implies a reduction of matching specificity (i.e., an increased probability of matching unrelated records). The question is whether we would privilege matching sensitivity or specificity. In my opinion RL might be applied

for studies aimed at estimating the frequency of a given disease or condition (e.g., prevalence, incidence, attack rate). In fact, in this case we would be able to capture all patients who present that disease, even if some false positives are necessarily included among cases captured by probabilistic RL. On the other hand, specificity would be privileged in investigating the association between healthcare and a given outcome, even if not all patients who experience the outcome are identified. In this case we would be able to ensure that patients classified as experiencing the outcome are true positives. Deterministic RL ensures this objective, since it is unlikely that said procedure may introduce false matching.

4. CONFOUNDING AND BEYOND

Misclassification is not the only type of bias researches are faced with. Confounding is also involved. Confounding occurs when the exposure-outcome relationship estimate is biased by the effect of one or several confounders. A confounder is an independent (causal) risk factor for the outcome analysed, and it is associated with the analysed exposure in the population, but it is not an intermediate step in the causal pathway between exposure and outcome [83, 84]. For example, low-density lipoprotein level would be expected to be a confounder in a study on the effects of statin treatment on the risk of cardiovascular events. The low-density lipoprotein level may lead a physician to prescribe a statin, and it may also be an independent risk factor for cardiovascular events.

4.1. Sources of confounding

Patient exposure to a medical therapy is determined by healthcare system-, physician-, and patient-level factors that may interact in complex and poorly understood ways [12, 16]. For example, a physician's decision regarding treatment may be based on an evaluation of the patient's health status and prognosis. Patients may initiate and comply with a new therapeutic regimen because of their disease risk and the benefits of treatment. Treatment initiation and adherence may also depend on a patient's physical and cognitive abilities. Patient and physician factors that determine the use of a treatment may directly affect health outcomes, or be related to them through indirect pathways. Several sources of bias can result from this process.

Confounding by indication or severity. A common, pernicious and often intractable form of confounding, endemic to pharmacoepidemiologic and healthcare research studies, is confounding by indication of treatment, that is physicians' tendency to prescribe medications to and perform procedures on patients who are most likely to benefit by them [16]. Because it is often difficult to assess medical indications and underlying disease severity and prognosis, confounding by indication often makes medications appear to cause outcomes they are meant to prevent [85, 86]. For example, statins, lipid-lowering drugs, reduce the risk of cardiovascular (CV) events in patients with CV risk factors. Hence, these drugs tend to be prescribed to patients who are perceived as presenting a higher CV risk. Incomplete control of CV risk factors can make statins appear to cause rather than prevent CV events [16].

Confounding by contraindication. When an adverse event is known to be associated with a therapy, confounding by contraindication is possible. For instance, women with a family history of venous thrombosis may avoid postmenopausal hormone therapy [87].

Confounding by functional status. Patients who are functionally impaired (defined as having difficulty performing daily living activities) may be less able to visit a physician or pharmacy and, therefore, may be less likely to collect prescriptions and receive healthcare services [16]. This phenomenon could exaggerate the benefits of prescription medications, vaccines, and screening tests. For example, functional status appeared to be a strong confounder in studies on both the effect of non-steroidal anti-inflammatory drugs (NSAIDs) and the influenza vaccine on all-cause mortality in the elderly [88-90].

Confounding by cognitive impairment. A similar form of confounding could result from differences in cognitive functioning. Depression may be considered as an example of such a bias because of the evidence that depression is a strong risk factor for several outcomes (such as CV disease [91]) as well as reduction in healthcare utilisation (e.g., lower compliance to treatment [92]).

The healthy user and healthy adherer

bias. Patients who initiate a preventive medication may be more likely than others to engage in other healthy, prevention-oriented behaviours [16]. For example, patients who start a preventive medication may be more likely to seek out preventive healthcare services, exercise regularly, moderate their alcohol consumption, and avoid unsafe and unhealthy activities. Incomplete adjustment to such behaviours can make use of preventive medications, spuriously associating them with a reduced risk of a wide range of adverse health outcomes.

Similarly, patients who adhere to treatment may also be more likely to engage in other healthy behaviours [93, 94]. Strong evidence of this "healthy adherer" effect comes from a meta-analysis of randomised controlled trials in which adherence to placebo was found to be associated with a reduced mortality risk [95]. This is clearly not an effect of the placebo but is rather due to the characteristics of patients who take a medication as prescribed. The healthy adherer bias is also evident in studies that reported associations between statin adherence and an increased use of preventive healthcare services, as well as a decreased risk of accidents [96, 97].

The healthcare access bias. Patients may vary substantially in their ability to access healthcare [16]. For example, patients who live in rural areas may have to drive long distances to receive specialised care. Other obstacles to accessing healthcare include cultural factors (e.g., trust in medical system), economic factors (e.g., affordability), immigration status, and institutional factors (e.g., restrictive formularies), all of which may have some direct or indirect effects on study outcomes.

4.2. Strategies to account for confounding: a general guide

It has been stated that confounding by indication, as well as other sources of confounding, are not an insurmountable problem [98]. This belief is based on two assumptions. The first is that the magnitude of confounding may be small because the treatment decisions of physicians may be poorly related to the pre-treatment prognostic characteristics of patients. Evidence to underpin this assumption comes from studies on the phenomenon of practice variation [99-103]. The reported magnitude of practice variation seems so large that some researchers have inferred that it could not arise from variability in patient characteristics (e.g., illness rates, insurance coverage, or preferences) [104] and, therefore, physicians pay little attention to individual patients' clinical characteristics when making decisions [105]. This assumption, however, does not take into account the fact that an unbalance in measured features among drug user categories has been repeatedly reported. For example, with the aim of avoiding falls and hip fractures, newer sedative-hypnotics are preferentially prescribed to frail elderly patients who more likely experience such outcomes. Frailty is difficult to measure in HCU databases [106], and has led to an overestimation of the association of newer sedative hypnotics with hip fractures, when compared with users of traditional benzodiazepines or non-users [107]. Generalising, since both magnitude and direction of confounding consequences are often unpredictable, the use of adequate tools to deal with confounding is always needed.

The second assumption is that HCU databases contain sufficient and sufficiently accurate information about pre-treatment prognostic differences between patients in order to make effective corrections by taking said differences into account. For example, Roos et al stated that "*administrative data has been shown to do nearly as well for risk adjustment as data that rely on physiological measures and physician judgment of health status*" [108]. All researchers, however, have not been comfortable with the role of observational studies using HCU databases [109-111]. Controversy focused on accuracy and sufficiency of information available in HCU databases to check differences in populations receiving different treatments and, in a broad sense, receiving care from different hospitals, providers, or healthcare systems.

Strategies to adjust for confounding vary depending on whether the potential confounders are measured in a given database. If confounders are measured, then usual (basic) strategies to account for confounding include those concerning study design (e.g., restriction and matching), and those concerning data analysis (standardisation, stratification and regression). These techniques, listed in Figure 4, are well described in standard epidemiology texts [112] and can be directly applied to database studies with the usual caveats.

However, the degree to which these clever devices achieve the goal of fully controlling confounding remains unpredictable, since unknown, unmeasured, or immeasurable confounders may strongly impact the findings. Residual confounding refers to factors that have been incompletely controlled, so that confounding effects of these factors may remain in the observed treatment-outcome effect. As depicted in Figure 4, strategies of accounting for residual confounding include those concerning study design (e.g., case-only designs) and data analysis (e.g., sensitivity analysis and instrumental variables approach).

The implications of these techniques when they are applied to HCU based studies need to be discussed. The purpose of the following paragraphs is to show different techniques, both usual and emerging, to adjust for measured and unmeasured confounders in the framework of observational studies based on HCU data.

4.3. Accounting for confounders through study design

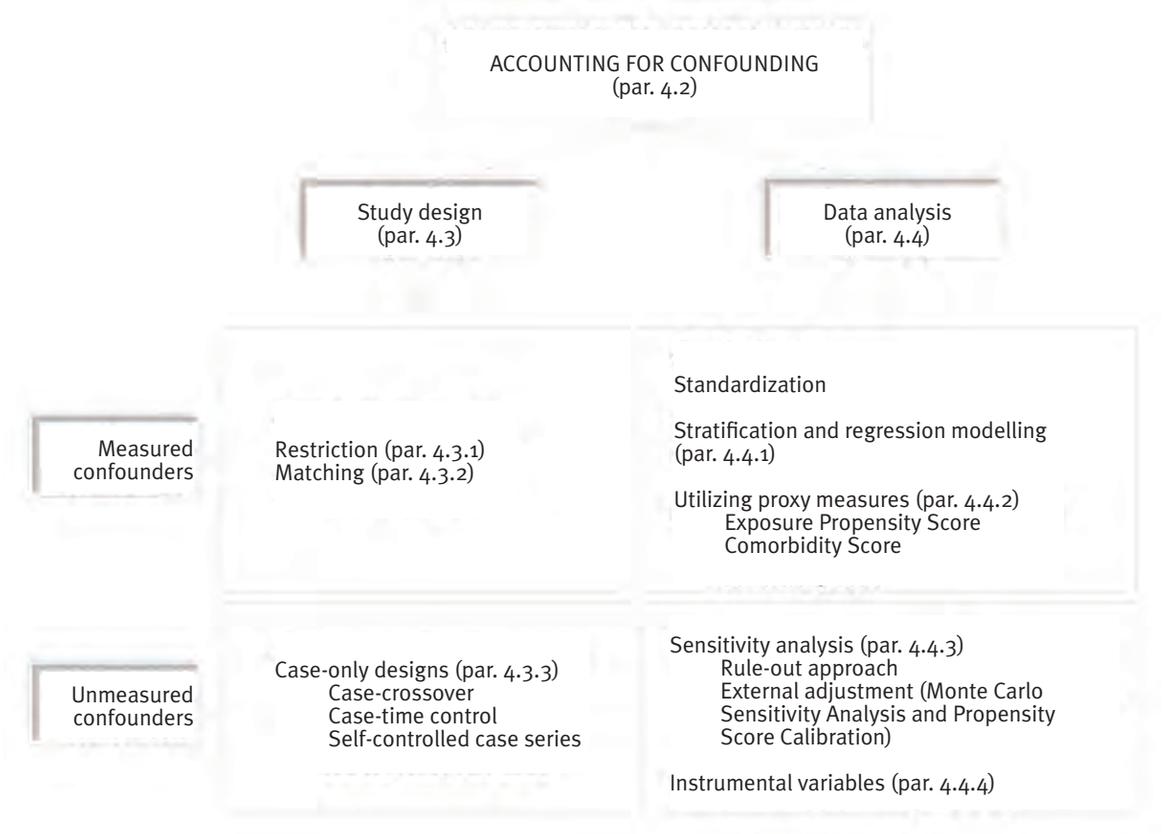
4.3.1. Restricting the study cohort

The basic idea of restricting a study cohort is to make its population more homogeneous regarding measured patient characteristics. Restriction will reduce the cohort size but population based HCU databases are so large that some restriction to improve the validity of findings will usually not impair precision significantly.

There are many different approaches to restriction in specific studies [113] and it is, therefore, difficult to provide general advice that fits specific study designs. However, several guiding principles can be identified that should be considered in HCU database studies on effectiveness and safety of medical interventions [114]. Three restrictions are generally worth considering in comparative effectiveness research [9].

FIGURE 4

STRATEGIES TO CHECK FOR MEASURED AND UNMEASURED CONFOUNDERS BY MEANS OF STUDY DESIGN AND DATA ANALYSIS



Source: Schneeweiss S [174], modified

Restricting to users and choosing a comparison group. Choosing a comparison group is a complex and, at times, subjective issue. The ideal comparison should comprise patients with identical distributions of measured and unmeasured risk factors of the study outcome [13]. Selecting comparison drugs that have the same perceived medical indication for head-to-head comparisons of active drugs (i.e., comparing the effects of two active and competing therapies that are prescribed under the assumption of identical effectiveness and safety [6]) will reduce confounding by selecting patients with the same indication [13]. Hence, excluding non-users and, consequently, actively comparing patients exposed to drugs with the same indication (i.e., the basic approach of comparative effectiveness research) has become increasingly popular in the field of observational studies that use HCU data [11]. However, new competitors within a class are often marketed for better efficacy, slightly expanded indications, or better safety, thus influencing the prescribing decisions of physicians [115]. New sources of confounding by indication can, therefore, arise.

Restricting to new users. As mentioned above, the basic cohort design identifies all patients in a defined population who were treated with the study medication during a defined study period. Such a cohort will consist of prevalent (ongoing) and incident (new) drug users; depending on the average duration (chronicity) of use, such cohorts may be composed predominantly of prevalent users and few new users. The estimated average treatment effect will, therefore, underemphasise effects related to drug initiation and will overemphasise the effects of long term use [18]. Furthermore, prevalent users of a drug have persisted by definition in their drug use, which may correlate with higher educational status and health-seeking behaviour, particularly if the study drug is treating a non-symptomatic condition, e.g., blood pressure (BP) lowering agents for treatment of hypertension, or statins for treatment of hyperlipidaemia [106, 116]. Consequently, the restriction to new initiators of the study drug (inception cohort) will mitigate those issues and will also ensure that patient characteristics are assessed before initiating treatment with the study drug.

The decision to rule out prevalent users from observational studies is underpinned by a recent study that evaluated the effect of excluding prevalent users of statins from a meta-analysis of observational studies on subjects with CV disease [117]. The pooled, multivariate adjusted mortality hazard ratio for statin use was 0.77 in 4 studies that compared incident users with non-users, 0.70 in 13 studies that compared a combination of prevalent and incident users with non-users, and 0.54 in 13 studies that compared prevalent users with non-users. The corresponding hazard ratio from 18 RCTs was 0.84. It appears that the greater the proportion of prevalent statin users in observational studies, the larger the discrepancy between observational and randomised estimates. The advantages of the new user design have been summarised by Ray [17].

Restricting to adherent patients. Patients dropping out of RCTs for reasons related to the study drug may cause bias. Non-informative discontinuation causes bias toward the null in intention-to-treat analyses. Physicians and regulatory agencies accept such bias because its direction is known and trial results are considered conservative regarding the drug's efficacy period. Discontinuation of treatment may also be associated with study outcomes through lack of perceived treatment effect and drug intolerance.

RCTs try to minimise bias from non-adherence by frequently reminding patients and by run in phases before randomisation to identify and exclude non-adherent patients. Adherence to drugs is substantially lower in routine care than in RCTs. It has been recently shown that in 36-37% of patients who start therapy for hypertension, hyperlipidaemia or type 2 diabetes, initial treatment is not followed by another specific prescription [118]. The study also showed that patients for whom an isolated prescription was issued presented clinical features (e.g., co-treatments and comorbidities), as well as a rate of hospitalisation for CV events, that was intermediate between those of patients for whom the considered medicaments were more regularly prescribed and those of individuals to whom such medicaments were never dispensed. Therefore, isolated users would be considered a heterogeneous category of individuals including those who would have needed continuous drug therapy and those for whom the lack of prescription renewal may be considered a later correction of inappropriate initial drug treatment. Similarly, it has been consistently shown that only 45%, 50% and 40% of patients who respectively

start therapy for hypertension, hyperlipidaemia or type 2 diabetes, refill their prescriptions after one year [119-121].

Several studies based on HCU data start follow-up after the second or third refill of the study drug in new user cohorts with the aim of excluding patients presenting less adherence. External validity (generalizability) is a matter of concern for this restricting criterion. However, in order to make a prescribing decision, physicians must assume that patients will take a drug as directed. If clinicians knew beforehand that a patient would not take a prescribed medication, they would not evaluate the appropriateness of the drug in the first place [9].

4.3.2. Matching

Matching is one of the techniques used to avoid confounding through study design. In a cohort study this is done by ensuring that once an exposed subject is enrolled for the study on a given date (e.g., because on that day he/she experienced the first prescription of the considered drug), one or more individuals belonging to the same population as the one that generated exposed ones are included on two conditions: (i) they did not have experienced exposure up to that date, (ii) regarding the exposed subject, they presented the same features which we think may confound the analysed association. A good example of a matched cohort study was presented by Ludvigsonn et al [122] who investigated the association between celiac disease (CD) and risk of renal disease. In this study, 14 336 CD patients and 69 875 patients without CD were matched by gender, age, calendar year, and country. As a result of matching, these variables had an equal distribution among both groups; therefore, these variables had no effect as confounders. Greenland & Morgenstern showed that matching can reduce the efficiency of a cohort study, even when it produces no sample-size reduction and even if the matching variable is a confounder [123]. This perhaps explains why this technique is not as popular in the field of observational cohort studies [124].

In a nested case-control study, a case (i.e., a subject who had experienced the outcome at a given date), is always matched with one or more controls (i.e., individuals who did not experience the outcome from the cohort entry until that date). As already discussed in par. 2.2., this design requires matching to ensure that members of each case-control(s) set have the same observational time-window length. In addition, case and relevant controls might be matched for other variables/features, which we think might confound the analysed association. For example, taking into account a patient who entered the cohort on a specific date and was 40 years old on entrance, one or more controls can be included in the cohort on the same date, with the same age on entrance and presenting the risk of experiencing the outcome at index date. Hence, we ensure that case and controls are balanced in terms of duration of observational period, and also of age. This prevents age from confounding the investigated association.

Because of their easiness and applicability, nested case-control studies are often designed taking into account matching for confounding adjustment. However, this technique has at least two weaknesses. First, once matching has been performed, the effect of matching variables on the outcome risk cannot be measured. Second, overmatching is always a hazard when this technique is used [125, 126]. This phenomenon can be explained with a theoretical example. Suppose we were able to match for all the variables affecting outcome. Both cases and controls would become almost completely similar, resulting in an odds ratio of approximately 1 [127]. In a broad sense, overmatching is introduced when at least one of these three conditions occurs: (i) the matching variable is not an independent risk factor for the outcome, thus reducing the efficiency of the analysis [128]; (ii) the matching variable is not associated with exposure (or is a proxy of exposure), thus resulting in an obscured exposure-outcome relation; (iii) the matching variable is on the causal exposure-outcome pathway (see par. 4.5.2). A motivating example is given by studies investigating the impact of gestation length and plurality on short-term outcome of in vitro fertilisation (IVF)-children. Since the high number of multiple and preterm births is an intrinsic part of current IVF practice, matching the control group by gestation length and/or the number of multiple births may yield misleading results on the total health impact of IVF, and should, therefore, be avoided [129].

Therefore, matching should be considered in case-control studies only if the matching

factor is known to be an independent risk factor for disease and unlikely to be on the causal path between the analysed exposure and the disease. However, given the scarce a priori knowledge on the possible effect of all analysed external variables, and owing to the large sample size of studies based on HCU data, other techniques would be better to count for confounders.

4.3.3. Case-only designs

Although cohort and nested case-control studies are widely accepted designs for the evaluation of the risks and benefits of post-licensure medications, these designs are vulnerable to confounding. In the late 1980s, alternative methods relying only on cases (i.e., without controls), termed case-only designs, were introduced with the aim of attempting to account for unmeasured confounders [130]. Case-only designs are attractive because cases are self-matched, which eliminates time-invariant confounders. They are generally less expensive, shorter in time, and simpler to carry out than conventional designs [131]. Among existing case-only designs, five have been used in pharmacoepidemiology: the case-crossover design [132], the self-controlled case series design, originally called case series analysis [133], the case-time-control design [134], the screening method [135] and the prescription sequence symmetry analysis also called the symmetry principle [136]. Of those designs, the first two are the most frequently used and will be briefly described below.

Case-crossover design. The case-crossover (CC) design was introduced by Maclure in 1991 to study the short-term effects of intermittent exposures on the risk of acute outcomes [132]. In a CC study, case-subjects are their own self-matched controls by using pre-defined time period(s). Probability of exposure is compared between a risk (or hazard) time-period immediately preceding the onset of the outcome of interest, and one, ore more than one, control period(s) preceding the risk period. Characteristics of time periods (width, number of control periods, gap from the onset the outcome, etc.) depends on the studied outcome and exposure. Only discordant pairs, i.e., cases exposed only in the risk period or only in the control period, contribute to the analysis. The odds ratio for the outcome is usually estimated by fitting conditional logistic regression model.

In a CC analysis, confounding by constant characteristics is implicitly eliminated. On the other hand, a bias due to exposure time trend is introduced. Case-time control (CTC) design adjusts the CC estimate for a time-trend in the exposure, by means of estimating it from a group of control subjects [134].

Self-controlled case series. The self-controlled case-series (SCCS) design was introduced by Farrington in 1995 to assess post-licensure adverse events related to vaccines, and more generally associations between acute outcome and transient exposure [133]. The SCCS design focuses on the relative incidence of a outcome between risk (post-exposure) and control time periods. Depending on the background knowledge, risk periods are defined during and/or after an exposure, when people are theorised to be at greater risk of the event. Control periods include all time periods with baseline risk, which may occur both before and after the exposure [137]. The model used to estimate the IRR is conditional Poisson regression; the condition is indeed the fact that all subjects are cases (all have experienced the analysed event). The tutorial by Whitaker et al [138] provides a very useful guide for this method.

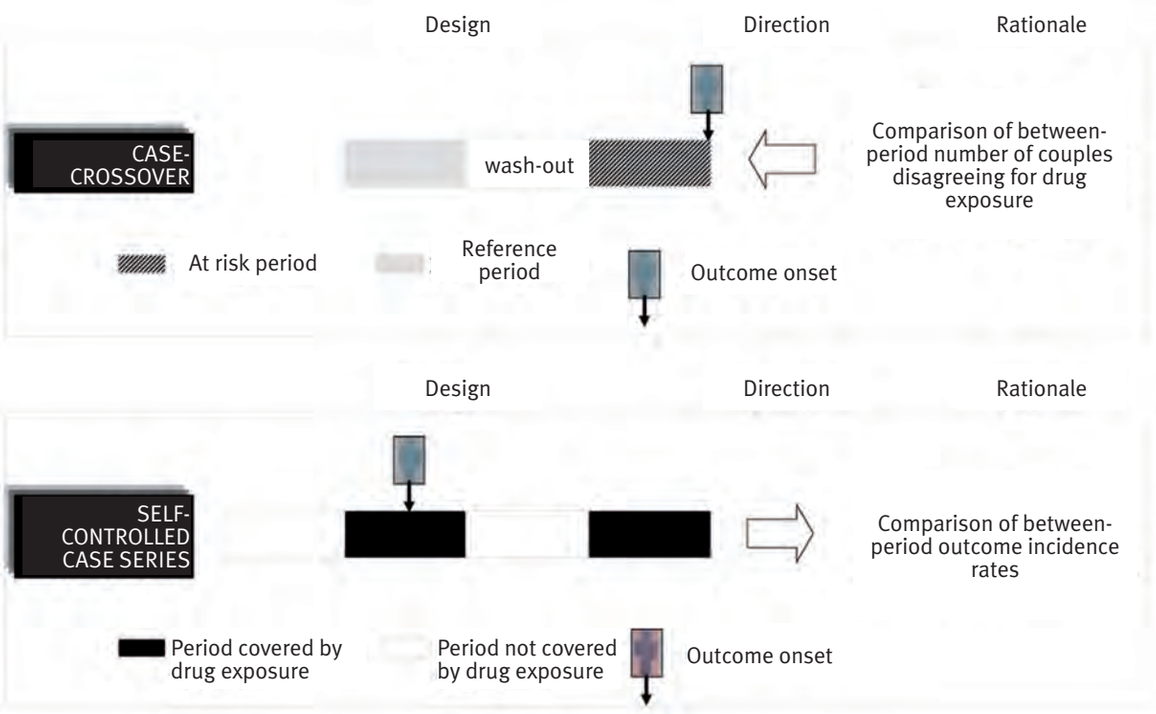
Figure 5 depicts these two main case-only designs. It clarifies that, while the CC design relates to a case control study (by comparing the probability of exposure between case and control periods), the SCCS relates to a cohort study (by comparing the probability of events between exposed and non-exposed periods). It should, however, be emphasised that both designs are particularly suited to study the effect of intermittent/transient exposures on the risk of acute outcomes [139]. Thus, to study chronic or cumulative effects of long term exposures, we must use other methods to account for residual confounding.

4.4. Accounting for confounding through data analysis

Causal graphs in Figure 6 illustrate data modelling techniques that account for confounding [9, 140]. Ideally, we would be able to fully assess all features that make unbalanced comparative groups and, then, to consider said features by means of stratification

FIGURE 5

SCHEMATIC REPRESENTATION OF CASE-ONLY DESIGNS (SELF-CONTROLLED CASE SERIES AND CASE-CROSSOVER)



and regression modelling (Figure 6, box A, and par. 4.4.1). However, as specified above, most non-randomised studies using HCU data with limited patient information will not be able to fully measure and adjust such confounders, and will, therefore, be unable to show the effect of exposure because of residual confounding. This difficulty to fully adjust to all possible confounders may be faced using several tools, such as resorting to proxy variables (Figure 6, box B, and par. 4.4.2), applying some sensitivity analysis techniques, e.g., external adjustment models (Figure 6, box C, and par. 4.4.3), or using instrumental variables. (Figure 6, box D, and par. 4.4.4).

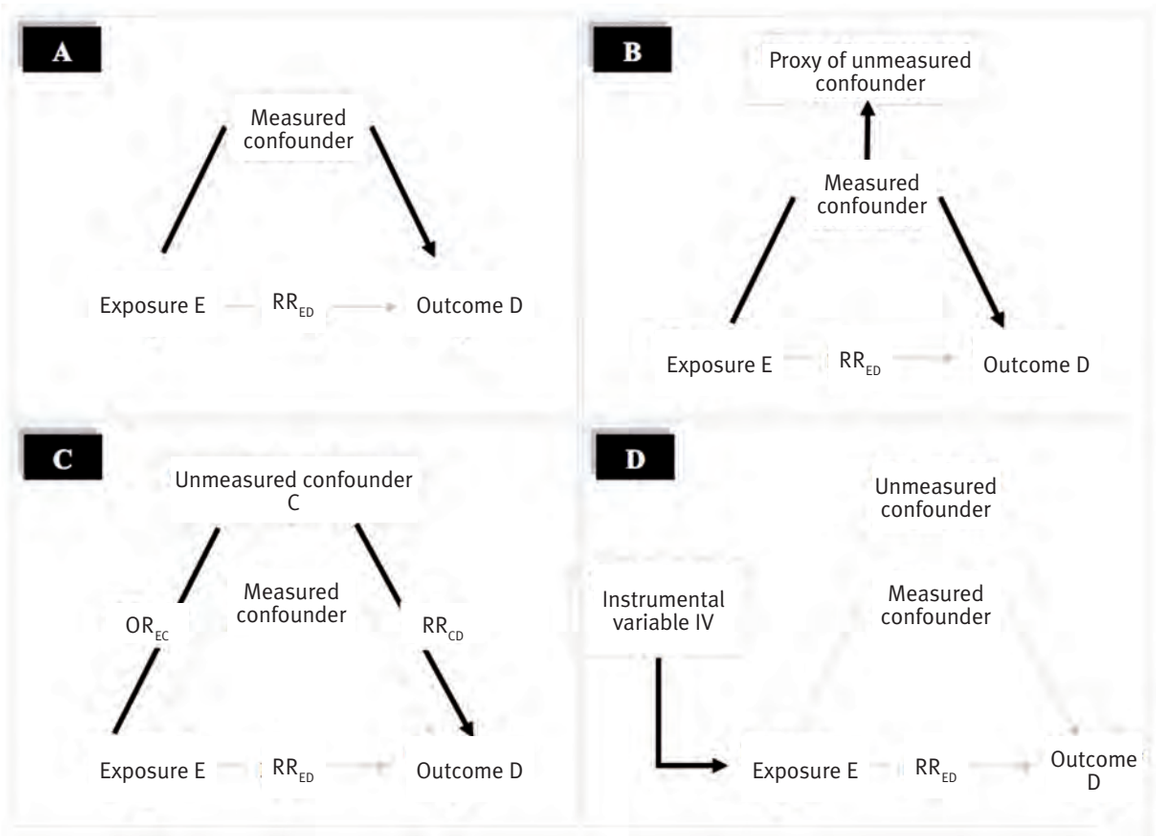
4.4.1. Stratification and regression modelling

Like restriction, stratification identifies patient subgroups based on measured patient factors [141]. In contrast to restriction, stratification does not discard the "unwanted" population but provides treatment effect estimates for all strata and combines them into one weighted summary effect measure [9]. In the absence of effect modification (e.g., treatment has the same effect in old and young patients) and under the assumption that all confounders were measured, stratified analyses will provide unbiased treatment effects. The large size of HCU databases permits many such subgroup analyses with substantial numbers of subjects and is an attractive alternative to wholesale restriction [9].

Regression analyses use statistical modelling to make stratified analyses more efficient [26]. Study design, conditioning the forms of study outcome, exposure of interest, and included covariates, will determine the regression model to be used. For cohort designs (see par. 2.1) that model time-to-event data with varying follow-up periods and censor study outcomes, the common analysis methodology is Cox proportional hazards regression. To be precise, this approach can easily handle exposures and covariates, whose values vary over time. When time-varying covariates are affected by time-varying treatment, marginal structural models may be required. A number of excellent textbooks describe how to analyse time-to-event data [142, 143]. For matched study

FIGURE 6

CAUSAL DIAGRAMS AND NOTATIONS SHOWING THE MECHANISMS OF CONFOUNDING AND APPROACHES TO CONTROL CONFOUNDING



Box A, confounding by unmeasured risk factors of the disease outcome (D) that are also associated with treatment exposure (E); **box B**, measured confounders as proxies for unmeasured confounders; **box C**, external adjustment using additional information on previously unmeasured risk factors (C); **box D**, an instrumental variable (IV) as an unconfounded substitute for treatment exposure (E).

The points on the graph representing the variables (exposure, disease, confounder, instrumental variable) are called nodes. The edges of the graph connecting any two nodes indicate relationships between these variables. An edge with an arrow connects causal variables to their effects. This, of course, signifies prior knowledge about how the variables operate in the population. A non-directional arc (an arc without arrowheads) is used to indicate that two variables are associated for reasons other than sharing an ancestor or affecting one another.

Sources: Schneeweiss S [9] and Greenland S et al [140], modified

designs (e.g., nested case-control design, see par. 2.2), conditional logistic regression may be considered. Finally, if the study outcome is binary with fixed follow up and is rare, Poisson regression with robust standard errors can be used to estimate relative risks and obtain correct confidence intervals [144, 145].

There are a number of analysis options that must be considered, which depend on the study question and details of the study design. For example, repeated outcomes, such as asthma attacks leading to emergency room admissions, can multiply the apparent number of subjects, resulting in falsely narrow standard errors [6]. Generalised estimating equations (GEE) are a frequently-used approach to account for correlated data [146]. Events are thus used as the unit of analysis, and standard errors are corrected to correlate covariate information within subjects [6].

Multilevel (also known as hierarchical, or mixed effects) regression modelling is a suitable approach to handle patient clustering with healthcare providers [147-150]. Failure

to use analytical methods that account for clustering can result in misleading conclusions. In a recent study, results from data consisting of acute myocardial infarction patients nested within treating physicians nested within hospitals with and without the use of multilevel modelling were contrasted [151]. The 95% CIs for hospital effects were much wider when multilevel logistic models were used compared with conventional logistic regression models that ignore clustering. Furthermore, substantially fewer statistically significant associations between patient outcomes and hospital characteristics were found when multilevel regression models compared with conventional regression models were used. When evaluating research using HCU data, readers need to carefully assess whether the statistical methods accounted for any clustering that may have been present in the data.

Approaches for such longitudinal data are described in detail in a number of textbooks [152, 153].

4.4.2. Using proxy measures

Researchers routinely adjust their analyses using proxy confounders. For example, the most common confounder of treatment effects is the patient's age. Although age itself does not cause outcomes, old age is associated with many conditions that may be incompletely recorded in the available data, but that are associated with the outcome and may be a determinant of pharmacologic interventions. Therefore, age can be considered an implicit proxy confounder [154].

In a study on the use of statins and CV outcomes, Seeger et al found that certain healthcare utilisation variables, such as frequency of lipid tests ordered and physician visits, were strong predictors of statin initiation and appeared also to be strong confounders [155]. In fact, although the frequency of lipid testing does not directly affect CV risk, it could be viewed as a proxy for concern about disease risk. Therefore, the frequency of tests may be associated with other risk modifying behaviours or with the underlying risk.

In several HCU databases, the number of proxies describing cross-sectional and longitudinal health status can quickly rise to several hundred, making it difficult to fit multivariate regression models for a limited number of observed outcomes even in large studies [156]. The most popular methods to efficiently adjust for a large number of proxies in database studies envisage constructing exposure propensity score and comorbidity score.

Exposure propensity score. In a cohort study, there are often substantial differences in the prevalence of measured patient factors between drug exposure groups that may lead to confounding, if these factors are also independent risk factors for the study outcome. Such factors need to be adjusted in further analyses. Instead of considering each factor individually, all patient characteristics can be combined into a single exposure propensity score (EPS), which is the estimated probability of treatment, given all covariates (i.e., the conditional probability of being treated given an individual's set of covariates [157, 158]), and is commonly calculated with logistic regression models.

The more formal EPS definition provided by Rosenbaum & Rubin [157] for the i^{th} subject (i = 1, . . . , N) is the conditional probability of assignment to a treatment (Z_i = 1) versus comparison (Z_i = 0) given observed covariates, x_i:

$$E(x_i) = pr(Z_i = 1 \mid X_i = x_i)$$

The underlying approach to propensity scoring uses observed covariates X to derive a "balancing score" $\beta(X)$ such that the conditional distribution of X given β (X) is the same for treated (Z = 1) and untreated patients (Z = 0) [157, 158]. There are three main applications of EPS: matching [158], stratification [159], and regression adjustment [160, 161]. When EPS is utilised in these standard applications, treatment effects are unbiased where measured covariates are nearly equally balanced across comparison groups [162, 163]. This transparent balancing of confounders promotes confidence in interpreting the results compared to other statistical modelling approaches [161]. This is why PS methodology has become increasingly popular to efficiently adjust large numbers of proxies in database studies.

Comorbidity score. Health status, as measured by disease history, has long been recognised as a major class of potential confounder in most observational studies. Over the last three decades, a variety of methods have been developed that might allow more uniform comorbidity adjustment across epidemiological studies. Six distinct comorbidity scores that

are useful for studies based on HCU data have been identified by a literature search and tested for their predictive performance [164]. Comorbidity measuring instruments include the Dartmouth-Manitoba method [165-167], the Chronic Disease Score [168], and its extended version [169], and the score proposed by Deyo et al [170], D'Hoore et al [171, 172], and Ghali et al [173]. All these scores reduce the number of covariates by summarising the diseases presented, that were recorded as ICD-9 code in a specific HCU database, in a single measure. Under the assumption that the score entirely captures information about the clinical profile of all included patients, which may be unrealistic in most practical settings, an analysis adjusting for the score produces exposure effect estimates unbiased by among-patients heterogeneity in clinical profile.

4.4.3. Sensitivity analysis

Basic sensitivity analyses of residual confounding attempts to understand how the strength of an unmeasured confounder and imbalance among drug exposure categories affects the observed or apparent *RR*. The observed exposure-outcome relative risk (*ARR*) can be expressed as the 'true' relative risk times a "bias factor," which expresses the imbalance of a binary confounding factor among exposed (P_{C1}) and unexposed subjects (P_{C0}) [174]:

$$ARR = RR \cdot \frac{P_{C1}(RR_{CD}-1)+1}{P_{C0}(RR_{CD}-1)+1}$$

where RR_{CD} is the strength of the association between confounder and disease outcome.

Schneeweiss [6, 9, 174, 175] describes two families of approaches investigating the impact of unmeasured confounding in the field of HCU databases: (1) identifying the strength of residual confounding that would be necessary to explain an observed drug-outcome association (rule-out approach); (2) external adjustment given additional information on single binary confounders from survey data using algebraic solutions and a Monte Carlo sampling procedure (Monte Carlo sensitivity analysis), or considering the joint distribution of multiple confounders from external sources of information (propensity score calibration).

The rule out approach. The approach aims at assessing the extent of confounding necessary to fully explain the observed findings, that is, when the observed point estimate would move to the null value. The hope is to rule out unmeasured possible confounders because they cannot possibly be strong enough to explain the observed association [175]. This approach was also called target-adjustment sensitivity analysis [176].

The approach consists in finding all combinations of OR_{EC} (i.e., the confounder-exposure odds ratio that measures the strength of the exposure-confounder association) and RR_{CD} (i.e., the confounder-outcome relative risk that measures the strength of the confounder-outcome association) that are necessary to move the observed point estimate of *ARR* to 1. It should be observed that OR_{EC} and RR_{CD} are respectively the left and right sides of the confounding triangle in Figure 6, box C [140].

Schneeweiss [174] showed that OR_{EC} is a function of the prevalence of the confounder among exposed (P_{C1}) and marginal probabilities of exposure P_E and confounder P_C:

$$OR_{EC} = \frac{P_{C1}(1-P_C-P_E+P_{C1})}{(P_C-P_{C1})(P_E-P_{C1})} \quad (1)$$

while, assuming no underlying true exposure-disease association (*RR*=1), Walker [177] showed that the apparent relative risk (A*RR*) is a function of P_{C1}, P_E, P_C, and the confounder-disease association RR_{CD}:

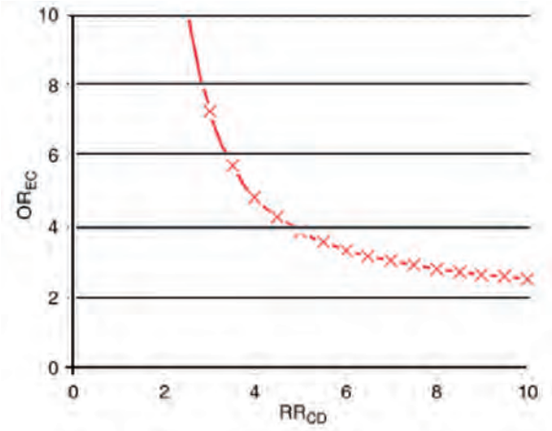
$$ARR = \frac{P_{C1}(RR_{CD}-1)+P_E}{(P_C-P_{C1})(RR_{CD}-1)-P_E+1}\frac{1-P_E}{P_E} \quad (2)$$

Since the primary interest is to explore the relationship between OR_{EC} and RR_{CD} for a given ARR, we need to solve Equation (2) for P_{C1} and substitute the derived term in Equation (1). Hence, we may calculate the pair (OR_{EC} - RR_{CD}) values that cancel the true exposure-disease association. Figure 7 provides an example of such a sensitivity analysis for residual confounding employed by Psaty et al [178] in a study on the association between calcium channel blocker use and acute MI.

The strength and usefulness of the rule-out approach is clear when little or nothing is known of possible confounders of the investigated association and their effect. However, when reliable additional data sources can be identified, external adjustment of the drug-outcome

FIGURE 7

EXAMPLE OF SENSITIVITY ANALYSIS OF RESIDUAL CONFOUNDING (RULE-OUT APPROACH)



This example by Psaty et al. [178] evaluates the effect of unmeasured confounders on the association between use of calcium channel blocker (CCB) and the risk of acute myocardial infarction (apparent relative risk or ARR = 1.57). Prevalence of the unobserved confounder of 0.2 and of CCB treatment of 0.01 were assumed from the study. Each line splits the area into two: the upper right area represents all parameter combinations of the strength of the associations between confounder and drug (OREC) and confounder and outcome (RRCD) that would create confounding by an unmeasured factor strong enough to move the point estimate of the ARR (ARR 1.57) to the null (ARR 1) or even lower, i.e., make the association go away. Conversely, the area to the lower left represents all parameter combinations that would not be able to move the ARR to the null

Source: Schneeweiss S [9], modified

association is always advisable [179].

External adjustment methods. If additional information is available, for example, a detailed survey in a sample of the main database study, alternative approaches to sensitivity analyses can be used to make adjustments for confounders that were not measured in the main study [174]. If internal validation studies are not feasible, or are too costly, external data sources can be used under certain assumptions. For example, structured electronic MR databases fed by general practitioners (GPs) operating in the same area, and covering a sample of the same population of the main database, may be used to measure a wide variety of characteristics that are not captured by HCU data, such as drug indications, lifestyle habits, body mass index, blood pressure measures and laboratory findings, among others. Thus, medical records can be used for external adjustment of unmeasured confounders in a variety of drug studies based on HCU data.

An example has been recently provided by Corrao et al [180, 181]. The authors observed that, compared with antihypertensive patients starting BP lowering therapy on a fixed-dose combination, those on an extemporaneous combination presented a 15% increase in CV risk. From a clinical point of view, this is rather puzzling because the reason why two antihypertensive drugs had different effects depending on whether they are administered in two distinct pills or in a single pill is unclear. Several uncontrolled factors may, however, influence the physician's decision to start therapy. For example, one can imagine that patients with more severe hypertension or worse clinical profile need a more aggressive therapy to quickly achieve BP control, and that this aggressive therapy is often obtained by dispensing an extemporaneous combination of two or more agents, rather than fixed combinations. Because severity of hypertension and clinical profile are independent predictors of the study outcome, failure to control them may lead to a confounding bias. Quantitative assessment of such a bias may provide more realistic estimates of the relationship between exposure and outcome.

The authors applied the following four-step procedure. First, they assessed the exposure-confounder association (that is, do patient's clinical characteristics affect the choice of prescribing a given antihypertensive drug

regimen?) by drawing out the corresponding data from an Italian network of general practitioners, the so called Health Search/ Cegedim Strategic Database. Second, some assumptions were made on confounder-outcome association (that is, do the clinical characteristics of patients affect the CV risk?). Third, these two types of external information were used to correct estimates generated from the main study consistently with Steenland & Greenland's approach [182]; they proposed estimating the bias factor by measuring the extent of the residual bias that would result from failure to check for a generic confounder:

$$Bias = \frac{\sum_{j=1}^{J} p_{j1} \cdot RR_j}{\sum_{j=1}^{J} p_{j0} \cdot RR_j} \quad (3)$$

where j indexes a generic confounder's category with $j = 0, 1$ for mild/moderate, severe hypertension; or $j = 0, 1, 2$ for an increasing number of chronic comorbidities. In equation (3) the risk ratio for the confounder-outcome association (RR_j) is weighted for the proportion of patients belonging to the same confounder category among those who started with a fixed-dose combination (p_{j0}) or with another therapeutic regimen (p_{j1}). The effect of starting with a given regimen for the CV risk, estimated from the main study, was separately adjusted for severity of hypertension and chronic comorbidities simply by dividing the original estimates by the bias factor (equation 3).

In the fourth step, the Monte-Carlo Sensitivity Analysis (MCSA), an expanded version of ordinary sensitivity analysis, was used with the aim of taking into account other random uncertainties of estimates obtained with external adjustment through a Monte-Carlo sampling procedure [182].

Figure 8 displays the CV odds ratios associated with the initial treatment regimen observed in HCU data (white squares) and after MCSA adjustment for severity of hypertension and CDS (black circles). Patients on an extemporaneous combination presented a higher CV risk than those on a fixed-dose combination. However, evidence that patients on an extemporaneous combination present a higher CV risk than those on a fixed-dose combination was cancelled after the adjustment for CDS, even if a relatively weak confounder - outcome association was imposed (scenario 2). This happens because of the large difference in clinical profile between patients starting BP-lowering therapy with a fixed-dose or extemporaneous combination. As a matter of fact, consistently with data from Health Search used for external adjustment, patients on an extemporaneous combination had higher prevalence of severe hypertension and worse clinical profile than those on a fixed-dose combination. Medical record data can be used to assess confounding bias unmeasured by HCU database with MCSA. The authors have supplied an SAS code that is useful for any application of this technique [181].

The main limitation of MCSA is that it is not helpful if several confounders are unmeasured and the joint effect of such confounders is unknown. However, external adjustment methods were recently extended to a multivariate adjustment for unmeasured confounders that use a new technique of propensity score calibration (PSC) [183]. In a validation study for each subject, the full database record is available along with detailed survey information. The goal is to compute an error-prone exposure PS within the validation population by only using database information, as well as an improved exposure propensity score that also includes survey information for each subject. The error component in the validation study is then quantified and can be used to correct the PS in the main study database by adopting established regression calibration techniques [71]. PSC implicitly takes into account the joint effect of unmeasured confounders that are measured only in the validation study, as well as relations between measured and unmeasured confounders. PSC can, therefore, elegantly overcome major limitations in the algebraic approach to external adjustment described above, though it may not perform well in situations that violate the surrogacy assumption of regression calibration [71, 184].

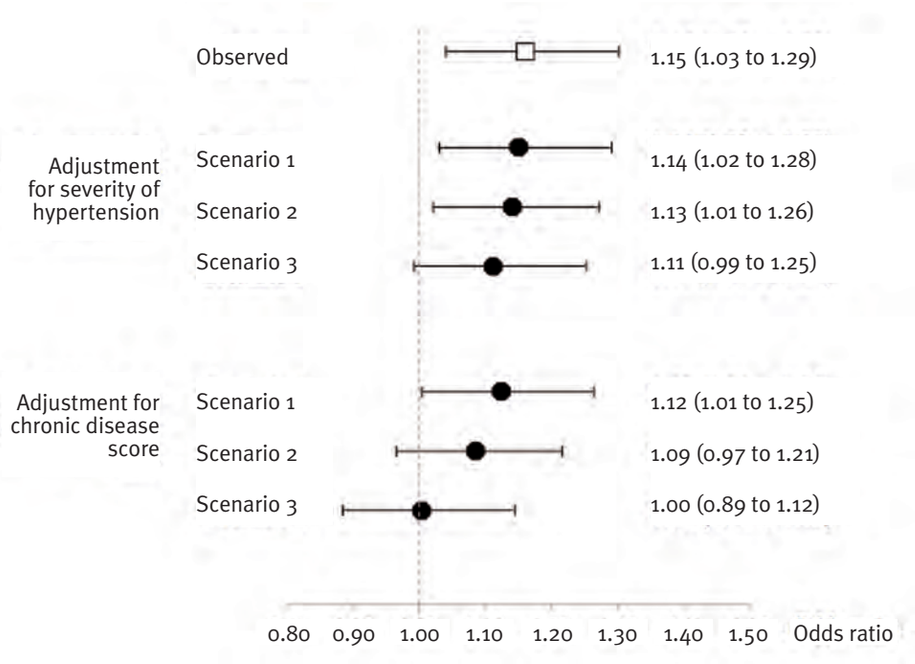
4.4.4. Instrumental variable estimation

To overcome the inability to check for residual confounding by unobserved factors, an analytic approach, known in econometrics as instrumental variable (IV) estimation [185], can provide unbiased estimates of causal effects in non-randomised studies [186] by mimicking random assignment of patients into groups of different likelihood for treatment [187].

A IV is a variable that is related to treatment but is unrelated to observed and unobserved

FIGURE 8

CHANGES IN ODDS RATIOS OF NON-FATAL CARDIOVASCULAR OUTCOMES ASSOCIATED WITH ANTIHYPERTENSIVE DRUG TREATMENT BASED ON AN EXTEMPORANEOUS COMBINATION OF TWO OR MORE DRUGS VERSUS A FIXED-DOSE COMBINATION (REFERENCE) AFTER EXTERNAL ADJUSTMENT FOR SEVERITY OF HYPERTENSION AND CHRONIC DISEASE SCORE



Adjustments were made with the Monte-Carlo Sensitivity Analysis taking into account (i) differences in clinical characteristics across strata of therapeutic regimens at entry (i.e., therapy with fixed-dose or extemporaneous combinations) estimated by means of external information (i.e., from Health Search medical records); (ii) three scenarios imposing that ln(RR) increase with a certain inclination across confounder categories

Source: Corrao G et al [181]

confounders. It is also unrelated to the outcome (Figure 6, box D), other than through treatment. Both characteristics are key assumptions for valid IV estimation. In the analysis, the unconfounded instrument substitutes the actual treatment status that may be confounded [9]. The instrument effect on the study outcome will be estimated and then rescaled by its correlation with the actual exposure [9]. The stronger the IV-treatment association is, the smaller the residual confounding effect will be. Moreover, precision of the IV estimation will improve [188].

IV estimation had not been used for the evaluation of drug effects until Brookhart et al [189] introduced physician prescribing preference as a promising tool for comparative effectiveness research. The basic idea is that there is a distribution of physician's preference for one drug over another that is largely independent of patient characteristics. One way to define a physician-prescribing preference tool is to categorise physicians into strong preferers of drug A if they prescribed it to 90% or more of their patients, whereas non-preferring doctors prescribe it in only 10% or less of cases. The variety of implementations of physician-prescribing preference is extensive, including the choice of a drug used by a physician for the most recent patient [189, 190]. In a study on the comparative effectiveness of selective COX-2 inhibitors versus non-selective NSAIDs, the last new NSAID prescription written by a physician was used to determine the IV value of the next patient. If the last patient received celecoxib, then for the next patient the physician is classified as a "celecoxib prescriber" [190]. This approach takes into account the fact that NSAID-prescribing preference may change within the study period. The analysis is performed with two-stage regression models adjusting

standard errors for correlations among patients clustering in the same physician's clinic [191].

4.5. Beyond confounding

4.5.1. Once again on the definition of confounding

The confounding hypothesis suggests that a third variable explains the relationship between exposure and outcome [192-196]. However, at least one definition of a confounder effect specifically requires that the third variable should not be an "intermediate" variable, as mediators are termed in epidemiological literature [197].

Consider a hypothetical example in which we are interested in assessing if periodontal disease (exposure) causes cardiovascular disease (CVD) (outcome) [198]. We also have measurements of C-reactive protein (CRP, i.e., the external third variable), a chronic inflammation marker that is associated with periodontal disease and CVD. In the univariate analysis, we find that there are statistically significant associations between periodontal disease and CVD, periodontal disease and CRP, and CRP and CVD. In the multivariate analysis, when we include CRP and periodontal disease in the same model to predict the CVD risk, we observe a null association between periodontal disease and CVD, and a positive association between CRP and CVD. According to Figure 9, box A, CRP elevation is a marker of a hyper-inflammatory trait, which causes both increased bone destruction in periodontal disease and atherosclerotic changes in CVD, but there is no actual association between periodontal disease and CVD. If we take this to be true, then we would conclude that periodontal disease does not cause CVD. The crude association we observed was through the backdoor path via CRP, which was closed when we adjusted for it in the multivariate analysis. According to Figure 9, box B, periodontal disease causes chronic low-grade infection, raises CRP levels and increases the risk of CVD. If we take this to be true, we would conclude that periodontal disease causes CVD only by a mediated effect via chronic inflammation (CRP). We can thus draw completely opposite conclusions with the same statistical data depending on which causal pathway we believe in.

This example clarifies that it does not suffice for a variable to be associated with both exposure and outcome to be considered a confounder. It is also necessary for the third external variable not to be an intermediate factor between exposure and outcome [199]. If this is the case, adjusting for the effect of the external variable (i.e., the intermediate variable or mediator) could substantially bias the estimated association between exposure and outcome.

This paragraph, which is based on intuitive approach, will enlarge on the inadequacy of conventional criteria to appropriately identify confounders, especially when overadjustment from mediation and collider stratification bias occurs [130, 200-203].

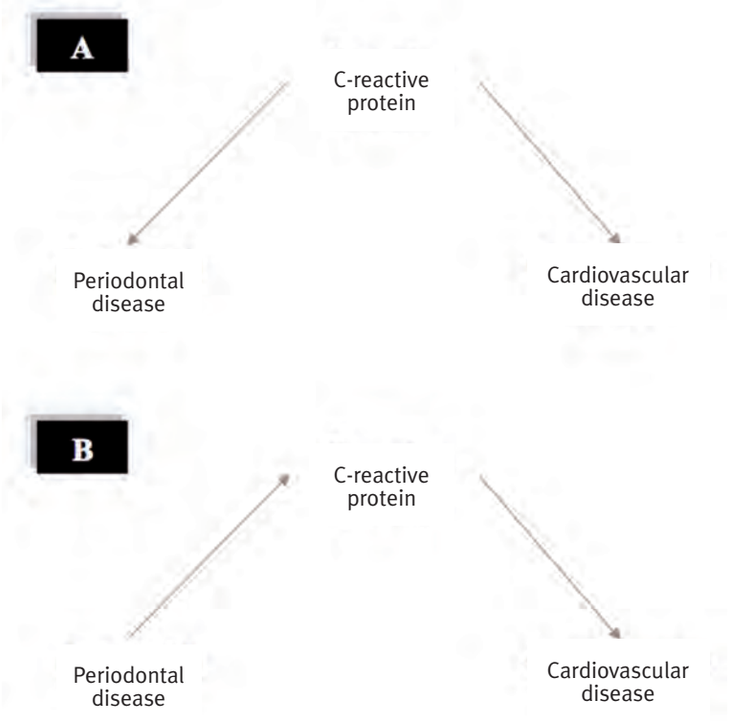
4.5.2. Intermediate variables and overadjustment

One reason why an investigator may begin to explore third variable effects is to elucidate the causal process by which exposure affects the outcome, i.e., a meditational hypothesis [192, 204]. When a meditational hypothesis is examined, the relationship between exposure and outcome is broken down into two causal paths [205]. One of these paths directly links exposure to the outcome (the direct effect), and the other links the independent variable to the dependent variable through a mediator (the indirect effect). An indirect or mediated effect implies that exposure causes the mediator (intermediate variable), which, in turn causes the outcome [206, 207].

An example has been recently provided by Roumie et al [208]. To study if incident use of oral antidiabetic drugs (OADs) is associated with the 12-month systolic BP level, and if this is mediated through body mass index (BMI), the authors included a cohort of veterans with hypertension who initiated metformin (n= 2 057) or sulfonylurea (n = 1 494) between 1 January 2000 and 31 December 2007. Figure 10 shows that sulfonylurea users had a 1.33 mmHg higher 12-month systolic BP than metformin users. However, when adjusting for BMI change, the difference in 12-month systolic BP between sulfonylurea and metformin users was not significant ($p = 0.72$), while one BMI unit change was associated with an increase in 12-month systolic BP of 1.07 mmHg ($p<0.0001$). These findings (i) strengthen the theory that the use of sulfonylurea increases systolic BP; (ii) suggest that the effect of OAD on BP change is likely

FIGURE 9

CAUSAL DIAGRAMS REPRESENTING THE EFFECT OF PERIODONTAL DISEASE ON CARDIOVASCULAR DISEASE



Box A, C-reactive protein causes periodontal disease on cardiovascular disease. **Box B**, C-reactive protein is an intermediate variable in the causal pathway

Source: Merchant AT & Pitiphat W [198]

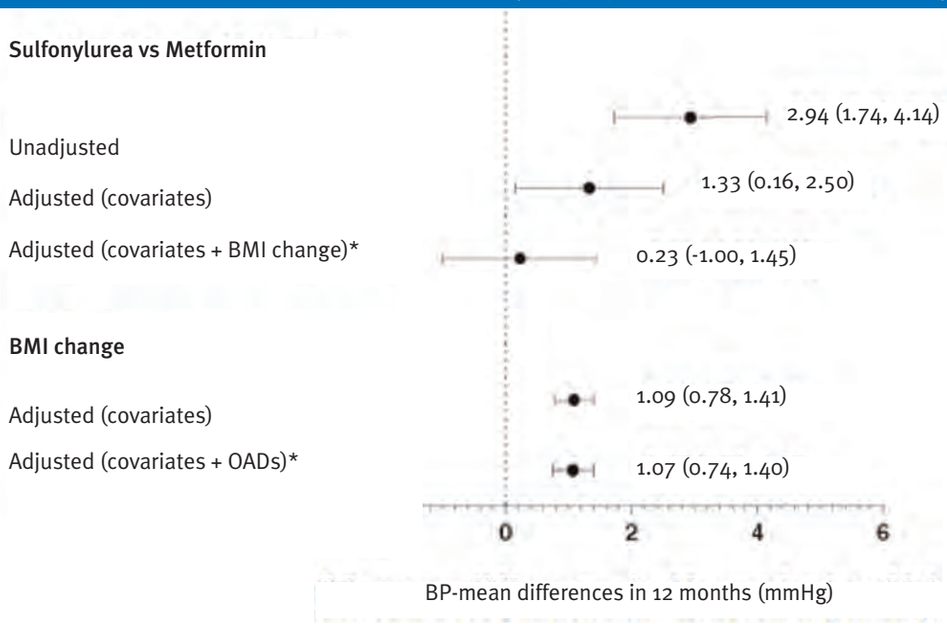
"mediated" by the beneficial effects of metformin in reducing BMI; (iii) demonstrate that a biased estimate of the effect of OAD on BP reduction is obtained when adjusting for the mediator (intermediate variable). In other words, the exposure-outcome association is obscured by the so called "overadjustment" effect [209].

Rothman & Greenland [112] discussed overadjustment in the context of intermediate variables. They clarified that, if checked during an analysis, intermediate variables would usually bias results towards null. Hence, instead of adjusting to exposure–disease associations, the assessment of mediation moves beyond the mere identification of exposure–disease associations toward an explanation of these relationships. The reasons for assessing mediation in epidemiology are compelling, and can be directly linked to extant mediational effects. Mediation analysis is very useful to open the "black box" between exposure and disease in epidemiologic studies [210].

Investigators are at times interested in separating total causal effect into direct and indirect effects, i.e., explaining the exposure-outcome directly and through the mediator, respectively. The goal of mediation analysis is to assess direct and indirect exposure on outcome effects. The conventional method of two-stage mediation analysis [211] involves fitting a series of linear regression models. Structural equation model (SEM) based methods have also been proposed for mediation analysis [212, 213]. In linear cases, it is straightforward to estimate mediation effects in the context of multiple stages by adopting the product of coefficients approach [214]. Some recent research has focused on mediation for binary or mixed types of variables [215-218]. Finally, a method that is applicable to multiple stages of mediation and mixed variable types using generalised linear models has been recently proposed [219].

FIGURE 10

MEAN DIFFERENCE IN 12-MONTH SYSTOLIC BLOOD PRESSURE (BP) AMONG HYPERTENSIVE DIABETIC PATIENTS INITIATED ON A NEW ORAL ANTIDIABETIC DRUG (OAD; I.E., SULFONYLUREA VS. METFORMIN)



*The effect of OAD was not adjusted for BP, only adjusted for covariates and adjusted for both covariates and body mass index (BMI) change. The effect of BMI change on BP was only adjusted for covariates and adjusted for both covariates and OAD

Source: Roumie CL et al [208]

4.5.3. Collider variables

Colliders are the result of two independent causes having a common effect [26]. When we include a common effect of two independent causes in our model, the previously independent causes become associated, thus opening a backdoor path between treatment and outcome. This phenomenon can be explained intuitively if we consider two causes for a lawn being wet (either the sprinklers have been turned on or it is raining). If we know that the lawn is wet and we know the value of one of the other variables (it is not raining), then we can predict the value of the other variable (the sprinkler must be on). Therefore, conditioning on a common effect induces an association between two previously independent causes, i.e., sprinklers being turn on and rain [26].

Consider a hypothetical study that uses HCU data to compare rates of acute liver failure between new users of CCB and diuretics [26]. As illustrated in Figure 11, CCB are mainly prescribed to older patients, while younger hypertensive subjects mainly receive diuretics (the age – antihypertensive arrow). On the other hand, older patients are more likely to receive treatment for erectile dysfunction (the age – treatment for erectile dysfunction arrow) and also have a long history of alcohol abuse (the age – alcohol abuse arrow). Finally, among the considered factors, only alcohol abuse truly causes acute liver failure. Nevertheless, antihypertensive treatment and liver disease should be associated when adjusting for treatment of erectile dysfunction.

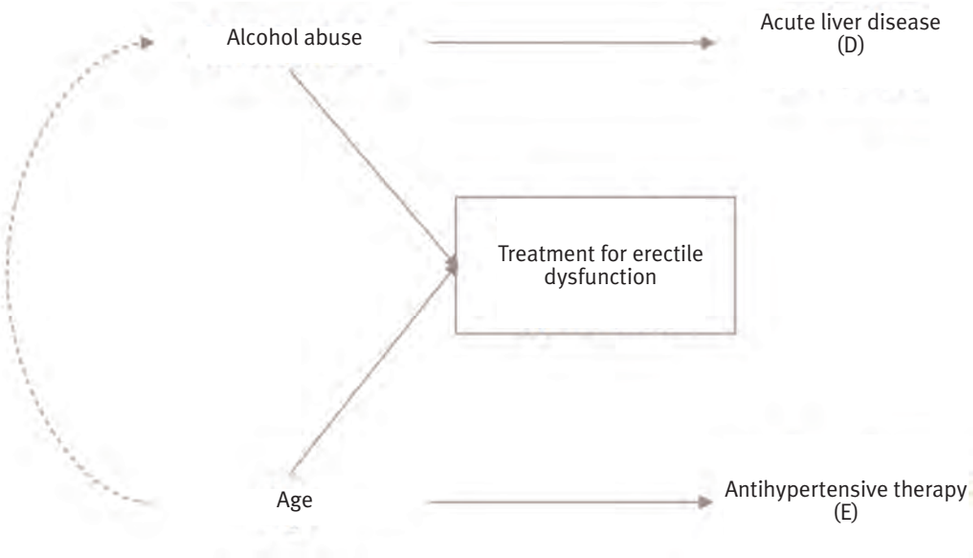
The introduced bias is known as collider stratification bias [220], or bias due to conditioning on a collider [221]. The term 'conditioning' refers to restriction, stratification or regression adjustment, which are the techniques described above to check for measured confounders [222].

5. RANDOM UNCERTAINTY, STATISTICAL SIGNIFICANCE AND CAUSALITY

Over the past decades the use of statistics in medical journals has increased remarkably. One consequence has been a shift in emphasis from basic results to undue focus on hypothesis

FIGURE 11

HYPOTHETICAL CAUSAL DIAGRAM ILLUSTRATING COLLIDER STRATIFICATION BIAS. AGE INFLUENCES TREATMENT WITH CCB (I.E., THE EXPOSURE VARIABLE OF INTEREST) AND TREATMENT FOR ERECTILE DYSFUNCTION



Unmeasured alcohol use influences impotence, erectile dysfunction treatment and acute liver disease (i.e., the outcome of interest). In this example antihypertensive treatment has no effect on liver disease, but antihypertensive treatment and liver disease would be associated when adjusting for medical treatment of erectile dysfunction. The box around erectile dysfunction treatment indicates adjustment and the conditional relationship is represented by the dotted arrow connecting age and alcohol use

Source: Agency for Healthcare Research and Quality [26], modified

testing. With this approach, data are examined in relation to a statistical "null" hypothesis, and practice has led to the mistaken belief that studies should aim at obtaining "statistical significance". Conversely, the purpose of most observational studies is to determine the magnitude of certain factor(s) in preventing the onset or in increasing the risk of a given outcome [223].

Overemphasis on hypothesis testing is particularly troubling in the setting of real-world data. Since the very large sample size and the wide number of possible associations may be simultaneously investigated, some caution is recommended when using and interpreting conventional statistical rationale.

5.1. Statistical significance and clinical relevance

The large sample sizes available in HCU databases have the potential to show statistical significance even when there are very small absolute differences. Although the conventional threshold for statistical significance of $p<0.05$ is widely used, one should keep in mind that it is arbitrary [224, 225].

The *p-value* for a regression model parameter results from testing the hypothesis that the measure of effect is null (e.g., it is equal to 1 if the measure is based on a ratio metric). The *p-value* is devoid of meaning with regard to the magnitude and clinical relevance of the observed effect, as it mirrors the precision of effect estimation. Excessive focus on a *p-value* of less than 0.05 can exaggerate the importance of statistically significant but clinically meaningless results. Likewise, this approach can discard potentially meaningful information gleaned from an analysis simply because the *p-value* exceeds an arbitrary threshold. By overemphasising the *p-value*, researchers may potentially distort the statistical model-building process by inappropriately adding or omitting certain variables, thereby resulting in suboptimal control of confounding and potentially invalid inferences. Rather than focusing on hypothesis tests (*p-values*), researchers should focus on estimates of effect

(point estimates and confidence intervals). In other words, estimation is preferable to tests of statistical significance [226]. Confidence intervals communicate both the strength of the relationship and the precision of the measure and are, therefore, more informative than point estimates accompanied by *p-values* [227].

A final point regarding *p-values* concerns the distinction between association and causality. Associations derived from observational data alone must not be construed to imply causality, regardless of the magnitude of the observed effect or its statistical significance. The strength of an association is only one of several factors that should be evaluated to establish causality [228]. Enhanced precision in the estimation of a measure of effect, as quantified by a low *p-value* or a narrow confidence interval, does not imply stronger evidence for causality [225].

5.2. Multiple comparisons

A wide number of possible associations may be investigated from observational studies based on HCU data. For example, we can compare the effect of different antihypertensive drugs in reducing the risk of CV outcomes (effectiveness) or the effect of different non-steroidal anti-inflammatory drugs in increasing the risk of serious gastrointestinal adverse events (safety). Simultaneous testing carries a type I error beyond the conventional threshold of α=0.05, and spurious conclusions may result. A type I error refers to the conclusion that a difference between two treatments exists when actually there is no difference [224]. All analyses of HCU data are essentially secondary data analyses and are, therefore, generally more susceptible to such statistical error. This implies a high probability of generating positive conclusions (i.e., statistical significant tests) simply by chance. On the other hand, the precision of estimates tends to vary considerably among the different exposures, depending on the number of persons to whom each drug is dispensed. Low precision estimates are more likely lead one to accept the null hypothesis of no association even when this is not true (i.e., low study power). When multiple effects are employed, the generation of false positive associations in addition to the lack of precision of some estimates, makes interpretation of the entire panel of results difficult. It would be helpful to minimise the total error in such hazard-surveillance programmes to clarify focus for further research [229].

A study investigating the association between astrological sign and the risk of hospitalisation for 223 of the most common diagnoses for hospitalisation has been recently published to illustrate how multiple hypotheses testing can produce not plausible associations [230]. Consistently with the common statistical criterion, the causes of hospitalisation occurring with a significantly higher probability compared to the remaining signs combined ($p<0.05$) were identified for each astrological sign. Of the (223 possible diagnoses • 12 astrological signs =) 2 676 potential associations, there were 72 causes of hospitalisation significantly associated with one or more specific astrological signs. For example, subjects born under Leo had a higher probability of gastrointestinal haemorrhage (P = 0.0447), while Sagittarians had a higher probability of humerus fracture (P = 0.0123), compared to all other signs combined!

It has been suggested that, when the researcher performs multiple comparisons using the same data, an adjustment should be made to maintain the experiment's α error at the prespecified level. Conventional multiple comparisons procedures, such as *Scheffé* or *Bonferroni* adjustments, may be useful for said purpose. For example, after the *Bonferroni* correction, none of the 72 associations found in the study on the association between astrological sign and cause of hospitalisation would have been significant [230]. Nevertheless, conventional multiple comparisons procedures have been described as unnecessary and ill-advised because they assume a global null hypothesis, which is neither plausible nor of interest, and because they may lead investigators to ignore unexpected but important findings [231].

Several methods for overcoming the problem of multiple comparisons have been developed in recent years. Bender & Lange provide an overview of methods that can be adjusted for multiple testing in medical and epidemiological literature [232]. Two of these methods will be briefly described in the following two sections. Additional details of statistical concepts and implementation procedures may be found elsewhere [233-235].

5.2.1. The False Discovery Rate method

In 1995 Benjamini & Hochberg proposed the False Discovery Rate (FDR) method which, primarily applied in the field of genetic research, controls the expected proportion of false rejections among all rejected theories [236].

FDR is the expected proportion of false signals (V) among those detected by the testing procedure (R), i.e., *E(V/R)*. The FDR controlling procedure is based on the adjustment of *p-values* concerning the *m* hypotheses to be tested simultaneously. Unfortunately, the estimation of *E(V/R)* is not straightforward. In practice, the use of the FDR method is simplified by implementation of an iterative algorithm based on a step-up procedure, which can be described as follows: (i) the original *p-values* are sorted from the larger to the smaller one; (ii) each adjusted *p-value* is calculated as the minimum value between the previous adjusted *p-value* and the original *p-value* multiplied by the ratio between the test number (*m*) and the number of *p-values* that still need to be adjusted (*m-j*). The iterative procedure to calculate the adjusted *p-value* as proposed by Benjamini & Hochberg in their original paper is [236]:

$$\tilde{p}_{(m)} = p_{(m)}$$
$$\tilde{p}_{(m-1)} = \min\left(\tilde{p}_{(m)}, \frac{m}{m-1} p_{m-1}\right)$$
$$\ldots$$
$$\tilde{p}_{(m-j)} = \min\left(\tilde{p}_{(m-j+1)}, \frac{m}{m-j} p_{m-j}\right)$$
$$\ldots$$

It is easy to see that the adjusted p-values are equal to or greater than the corresponding original ones, and that the number of signals is consequently reduced. However, when the number of comparisons becomes very large, only the smallest adjusted *p-values* are greater than the originals.

5.2.2. Empirical-Bayes methods

It has been repeatedly proven that, when applied properly, Empirical-Bayes (EB) methods dramatically outperform conventional methods when one wishes to obtain multiple predictions or estimates that minimise the total error [237-241]. EB adjustment is useful under the following circumstances: (i) a large number of comparisons are made; (ii) the comparisons can be grouped into sets within which all comparisons can be considered similar or exchangeable; (iii) random error is present and presumably accounts for much of the observed variation in the parameters estimated; (iv) investigators must choose which comparisons to investigate further; and (v) there is a significant cost associated with such additional investigations [242].

The basic idea of EB adjustments for multiple associations is that the observed spread of the estimated effects around their mean is larger than the variation of the true but unknown effects. EB methods attempt to estimate this extra variation from data and then use said estimate to adjust the observed effects. Typically, this adjustment shrinks outlying effects towards their mean, especially if the estimate to be adjusted has a large individual variance. A consequence of this shrinkage is that the overall variance of the EB-adjusted estimates is smaller than that of the unadjusted estimates. EB estimators belong to a class of shrinkage estimators with a long history in statistical literature [243, 244]. This class includes estimators based on hierarchical, multilevel, and mixed models [245]. Simulation and empirical studies have shown that EB adjustment can provide more accurate point estimates and narrower confidence intervals than the original estimates [234, 236, 239, 242, 244, 246].

EB adjustment is now being applied in several fields of epidemiologic research (e.g., in occupational [234, 238, 242], genetic [246], and nutritional [247] epidemiology). However, to our best knowledge, such an approach has rarely been used in comparative effectiveness studies based upon HCU data [248].

6. RECOMMENDATIONS FOR GOOD RESEARCH PRACTICE

When making healthcare decisions, patients, healthcare providers, and policymakers routinely seek unbiased information about the effects of treatment on a variety of health outcomes. High quality research can reduce uncertainty about the net benefits of treatment by providing scientific evidence and other objective information to inform healthcare decisions. Nonetheless, it is estimated that more than half the medical treatments lack valid evidence of

effectiveness [249, 250] particularly for long-term outcomes. Therapies that demonstrate efficacy in experimental settings, like randomised controlled trials, may perform differently in general clinical practice where there is a wider diversity of patients, providers, and healthcare delivery systems [251, 252]. The effects of these variations on treatment are sometimes unknown but can significantly influence the net benefits and risks of different therapy options in individual patients.

One of the most important elements in designing a study is the development of a study protocol, which is the project that guides and governs all aspects of how a study will be conducted. A study protocol manages the execution of a study to help ensure the validity of final study results. It also provides transparency regarding how the research is conducted and improves reproducibility and replication of the research by others, thereby potentially increasing the credibility and validity of a study's findings [26]. For studies designed as randomised clinical trials, study protocols are common and standards have been defined to establish the contents of these protocols. However, for other study designs, such as observational ones, there are few standards specifically for what elements are recommended to be included in a study protocol. As a result, there are a wide range of practices among investigators [253] in an attempt to summarise the current overview of methods adopted to overcome the drawbacks of an observational approach based on real-world data, this final paragraph provides a checklist for the development of a study protocol to increase the transparency of the rationale behind study design selection and to clearly define methods. Several references have inspired the following checklist, specifically the document entitled "*Developing a Protocol for Observational Comparative Effectiveness Research (OCER): A User's Guide*" by the Healthcare Research and Quality (AHRQ) [26], the "*Guidelines for good pharmacoepidemiology practices*" from the International Society of Pharmacoepidemiology [227], the task force report by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) concerning "*Good Research Practices for Retrospective Database Analysis*" [2, 13, 14, 65, 256], besides some key papers on this issue [11-13, 257].

6.1. Study objectives and questions (background)

- Describe the knowledge or information to be gained from the study
- Articulate the main study objectives in terms of a highly specific research question or set of related questions that the study will answer
- Synthesise the literature and characterise the known effects of exposures and interventions on patient outcomes

6.2. Data sources

- Propose data source(s) that include data required to address primary and secondary research questions
- Describe details of data source(s) selected for the study
- Describe validation or other quality assessments that have been conducted on the data source that are relevant to the data elements required for the study
- Describe what patient identifiers are necessary for the research purpose, how they will be protected, and permissions/waivers required
- Provide details on data linkage approach, and the quality/accuracy of linkage, if applicable

6.3. Study design

- Provide rationale for study design choice and describe key design features: e.g., cohort, nested case-control, case-cohort, case-crossover, case-time-control, self-controlled case series
- Define start of follow-up (baseline)
- Define inclusion and exclusion criteria at start of follow-up (baseline)
- Define exposure (treatments) of interest at start of and during the follow-up
    - Propose a definition of exposure that is consistent with the clinical/conceptual basis for the research question
    - Provide a rationale for exposure

time window choice
    - Describe the proposed data source(s) and explain how they are adequate and appropriate for defining exposure
    - Provide evidence of validity of the operational definition of exposure with estimates of sensitivity, specificity, and positive predictive value, when possible
    - State the direction of potential sources of differential and non-differential misclassification and how that could influence the acceptance or rejection of the null hypothesis
    - Propose strategies for reducing exposure misclassification
- Choose concurrent, active comparators from the same source population (or justify use of no treatment comparisons/historical comparators/different data sources)
    - Discuss potential bias associated with comparator choice and methods to minimise such bias, when possible
- Define outcome(s) of interest
    - Propose primary and secondary outcomes that directly correspond to research questions
    - Provide clear and objective definitions of clinical outcomes
    - Provide evidence of sensitivity, specificity, and positive predictive value of the outcome, when possible
    - Address issues of differential and non-differential misclassification related to the outcome and propose strategies for reducing bias, where possible
- Define key covariates and their potential for confounding (or other action on the relationship of interest)
    - Conduct a thorough literature review to identify all potential confounding factors that influence treatment selection and outcome. Create a table detailing the expected associations
    - When measuring comorbidity, select a measure that has been validated in a population most similar to the study and for the outcome under investigation, where possible
    - Provide information about data source(s) for exposure, outcome and key covariates, acknowledging the strengths and weaknesses of the data source for measuring each type of variable

6.4. Statistical issues

- Describe the key variables of interest with regard to factors that determine appropriate statistical analysis
    - Independent variables (when are they measured, fixed or time-varying; e.g., exposures, confounders, effect modifiers)
    - Dependent variables or outcomes (continuous or categorical, single or repeated measure, time-to-event)
    - State if there will be a "multi-level" analysis (e.g., looking at effects of both practice level and patient level characteristics on outcome)
- Propose descriptive analysis according to study groups (e.g., treated with the compared drugs)
    - Should include the available independent variables (e.g., exposure, confounders, effect modifiers, etc.…)
    - Conduct a stratified analysis prior to undertaking a more complex analysis because of its potential to provide important information on relevant covariates and how they could be optimally included in a model
- Propose the model that will be used for primary and secondary analysis objectives
    - The functional form of the model should take into account study objectives, study design (independent vs. dependent observations, matched, repeated measurement); nature of outcome measure (e.g., continuous, categorical, repeated measures, time to event), fixed and time-varying exposure and other covariates, assessment of effect modification/heterogeneity, censored data, and the degree of rarity of outcome and exposure
    - Include variables that are only weakly related to treatment selection because they may potentially reduce bias more than they increase variance

    - All factors that are theoretically related to outcome or treatment selection should be included despite statistical significance at traditional levels of significance
    - Check the assumptions of the model before fitting it; if assumptions are violated, alternative techniques should be implemented.
    - Performance measures (R2, area under ROC curve) should be reported and a qualitative assessment of these measures should be discussed regarding the explanation of variance or discrimination of the model in predicting outcome
    - Regression diagnostics including goodness of fit should be conducted and reported
    - Consider presenting the final regression model, not only the adjusted treatment effects. If journal or other publications limit the amount of information that can be presented, the complete regression should be made available to reviewers and readers in an appendix.
- Propose and describe planned sensitivity analyses
    - Consider the effect of changing exposure, outcome, confounder, or covariate definitions or classifications
    - Assess expected impact of unmeasured confounders on key measures of association
    - If sensitivity analyses are performed for different assumptions regarding the confounding structure, report directed acyclic graphs representing the assumptions of the respective sensitivity analyses
- Report considerations concerning the study size under several potential scenarios that vary the baseline risk of the outcome, the minimum clinically relevant treatment effect (i.e., the size of the smallest potential treatment effect that would be of clinical relevance), and the required power
- Justify the choice of a given magnitude of first type error, and propose and describe planned methods for overcoming the problem of multiple comparisons, when suitable
- In presenting and discussing the results, the greatest emphasis should be placed on bias and confounding, rather than on the role of chance
    - Confidence intervals, rather than p-values, should be relegated to a small part of both the results and discussion sections as an indication, but no more, of the possible influence of chance imbalance on the result

ACKNOWLEDGEMENTS: *Prof. Corrao thanks the Italian Ministry for University and Research ("Fondo d'Ateneo per la Ricerca," year 2011) for the support provided. A warm thank you to Colleagues, PhD students and young Collaborators at the Department of Statistics and quantitative methods, "Biostatistics, epidemiology and public health" Unit, for many helpful discussions and manuscript reviewing, and a special thank you to Prof. Antonella Zambon, Dr. Andrea Arfè, Dr. Arianna Ghirardi, Dr. Silvana Romio and Dr. Giulia Segafredo.*

References

[1] Motheral BR, Fairman KA. The use of claims databases for outcomes research: rationale, challenges, and strategies. Clin Ther 1997; 19: 346-66

[2] Berger ML, Mamdani M, Atkins D, et al. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report-Part I. Value in Health 2002; 12: 1044-52

[3] Dreyer NA, Schneeweiss S, McNeil BJ, et al. GRACE principles: recognizing high-quality observational studies of comparative effectiveness. Am J Manag Care 2010; 16: 467-71

[4] Tamblyn R, LaVoie G, Petrella L, Monette J. The use of prescription claims databases in

pharmacoepidemiological research: the accuracy and comprehensiveness of the prescription claims database in Quebec. J Clin Epidemiol 1995; 48: 999-1009

[5] Takahashi Y, Nishida Y, Asai S. Utilization of health care databases for pharmacoepidemiology. Eur J Clin Pharmacol 2012; 68: 123-9

[6] Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005; 58: 323-37

[7] Suissa S. Novel Approaches to Pharmacoepidemiology Study Design and Statistical Analysis. In: Pharmacoepidemiology (4th edn). Strom BL (ed). Wiley, New York, 2005; 811-29

[8] Lohr KN. Emerging Methods in Comparative Effectiveness and Safety. Symposium Overview and Summary. Med Care 2007; 45: S5-S8

[9] Schneeweiss S. Developments in post-marketing comparative effectiveness research. Clin Pharmacol Ther 2007; 82: 143-56

[10] Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects—advantages and disadvantages. Nature Clinical Practice Rheumatology 2007; 3:725-32

[11] Schneeweiss S, Gagne JJ, Glynn RJ, et al. Assessing the comparative effectiveness of newly marketed medications: methodological challenges and implications for drug development. Clin Pharmacol Ther 2011; 90: 777-90

[12] Sox HC, Goodman SN. The methods of comparative effectiveness research. Annu Rev Public Health 2012; 33: 425-45

[13] Cox E, Martin BC, Van Staa T, et al. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research. Practices for retrospective database analysis task force report-Part II. Value Health 2009; 12: 1053-61

[14] Johnson ML, Crown W, Martin BC, et al. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources. The International Society For Pharmacoeconomics and Outcomes Research Good Research. Practices for retrospective database analysis task force report-Part III. Value Health 2009; 12: 1062-73

[15] Freedman LS, Schatzkin A, Midthune D, et al. Dealing with dietary measurement error in nutritional cohort studies. J Natl Cancer Inst 2011; 103: 1086-92

[16] Brookhart MA, Sturmer T, Glynn RJ, et al. Confounding control in healthcare database research challenges and potential approaches. Med Care 2010; 48(suppl 1): S5-8

[17] Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 2003; 158: 915-20

[18] Lubin JH: Extensions of analytic methods for nested and population-based incident case-control studies. J Chronic Dis 1986; 39: 379-88

[19] White E, Hunt JR, Casso D: Exposure measurement in cohort studies: the challenges of prospective data collection. Epidemiol Rev 1998; 20: 43-56

[20] Fisher LD, Lin DY: Time-dependent covariates in the Cox proportional-hazards regression model. Annu Rev Public Health 1999; 20: 145-57

[21] Marubini E, Valsecchi MG. Analysing survival data from clinical trials and observational studies. New York: John Wiley & Sons; 1995

[22] Mantel N. Synthetic retrospective studies and related topics. Biometrics 1973; 29: 479-86

[23] Essebag V, Platt RW, Abrahamowicz M, et al. Comparison of nested case-control and survival analysis methodologies for analysis of time-dependent exposure. BMC Medical Research Methodology 2005; 5: 5, available from: http://www.biomedcentral.com/1471-2288/5/5

[24] Hutchinson GB. Leukemia in patients with cancer of the cervix uteri treated with radiation. J Natl Cancer Inst 1968; 40: 951-82

[25] Prentice RL. A case–cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986; 73: 1-11

[26] Agency for Healthcare Research and Quality (AHRQ). Developing a protocol for observational comparative effectiveness research (OCER): a user's guide. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/research-DRAFT-COPY_AllChapters.pdf

[27] Collet JP, Schaubel D, Hanley J, et al. Controlling confounding when studying large pharmacoepidemiologic databases: a case study of the two-stage sampling design. Epidemiology 1998; 9: 309-15

[28] Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol 1996; 144: 207-13

[29] O'Malley KJ, Cook KF, Price MD, et al. Measuring Diagnoses: ICD Code Accuracy. Health Serv Res 2005; 40: 1620-39

[30] Van Walraven C, Bennett C, Forster AJ. Administrative database research infrequently used validated diagnostic or procedural codes. J Clin Epidemiol 2011; 64: 1054-9

[31] Chyou PH. Patterns of bias due to differential misclassification by case-control status in a case-control study. Eur J Epidemiol 2007; 22: 7-17

[32] Copeland KT, Checkoway H, McMichael AJ, Holbrook

RH. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol 1977; 105: 488-95

[33] Pekkanen J, Sunyer J, Chinn S. Nondifferential disease misclassification may bias incidence risk ratios away from the null. J Clin Epidemiol 2006; 59: 281-9

[34] Slee VN, Slee D, Schmidt HJ. The tyranny of the diagnosis code. N C Med J 2005; 66: 331-7

[35] Hsia DC, Krushat WM, Fagan AB, et al. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. N Engl J Med 1988; 318: 352-5

[36] Cimino JJ. Formal descriptions and adaptive mechanisms for changes in controlled medical vocabularies. Methods Inf Med 1996; 35: 202-10

[37] Vardy DA, Gill RP, Israeli A. Coding medical information: classification versus nomenclature and implications to the Israeli medical system. J Med Syst 1998; 22: 203-10

[38] Feinstein AR. ICD, POR, and DRG. Unsolved scientific problems in the nosology of clinical medicine. Arch Intern Med 1988; 148: 2269-74

[39] Cimino JJ. An approach to coping with the annual changes in ICD9-CM. Methods Inf Med 1996; 35: 220

[40] Hogan WR, Slee VN. Measuring the Information Gain of Diagnosis vs. Diagnosis Category Coding. AMIA Annu Symp Proc 2010; 2010: 306-10

[41] Kelsey JL, Whittemore AS, Evans AS, Thompson WD. Methods in observational epidemiology. 2nd edition. New York: Oxford University Press; 1996

[42] Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol 2004; 57: 131-41

[43] Romano PS, Mark DH. Bias in the coding of hospital discharge data and its implications for quality assessment. Med Care 1994; 32: 81-90

[44] Antoniou T, Zagorski B, Loutfy MR, et al. Validation of case-finding algorithms derived from administrative data for identifying adults living with human immunodeficiency virus infection. PLoS One 2011; 6: e21748

[45] Hux JE, Ivis F, Flintoft V, et al. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. Diabetes Care 2002; 25: 512-6

[46] Corrao G, Botteri E, Bagnardi V, et al. Generating signals of drug-adverse effects from prescription databases and application to the risk of arrhythmia associated with antibacterials. Pharmacoepidemiol Drug Saf 2005; 14: 31-40

[47] Navarese EP, Buffon A, Andreotti F, et al. Meta-analysis of impact of different types and doses of statins on new-onset diabetes mellitus. Am J Cardiol 2013; doi:10.1016/j.amjcard.2012.12.037

[48] Delgado-Rodríguez M, Llorca J. Bias. J Epidemiol Community Health 2004; 58: 635-41

[49] Stergachis AS. Record linkage studies for postmarketing drug surveillance: data quality and validity considerations. Drug Intell Clin Pharm 1988; 22: 157-61

[50] Levy AR, O'Brien BJ, Sellors C, et al. Coding accuracy of administrative drug claims in the Ontario Drug Benefit database. Can J Clin Pharmacol 2003; 10: 67-71

[51] McKenzie DA, Semradek J, McFarland BH, et al. The validity of Medicaid pharmacy claims for estimating drug use among elderly nursing home residents: the Oregon experience. J Clin Epidemiol 2000; 53: 1248-57

[52] Strom BL. Overview of automated databases in pharmacoepidemiology. In: Strom BL, editor. In: Pharmacoepidemiology (4th edn). Strom BL (ed). Wiley, New York, 2005; 219-22

[53] West S, Savitz DA, Koch G, et al. Recall accuracy for prescription medications: self report compared with database information. Am J Epidemiol 1995; 142: 1103-12

[54] West S, Strom BL, Freundlich B, et al. Completeness of prescription recording in outpatients medical records from a health maintenance organization. J Clin Epidemiol 1994; 47: 165-71

[55] Peterson AM, Nau DP, Cramer JA, et al. A checklist for medication compliance and persistence studies using retrospective databases. Value in Health 2007; 10: 3-12

[56] WHO Collaborating Centre for Drug Statistics Methodology. ATC index with DDD. Oslo, Norway: WHO; 2003

[57] de Abajo FJ, Garcia Rodriguez LA. Risk of upper gastrointestinal bleeding and perforation associated with low-dose aspirin as plain and enteric-coated formulations. BMC Clin Pharmacol 2001; 1: doi:10.1186/472-6904-1-1

[58] Ilkhanoff L, Lewis JD, Hennessy S, et al. Potential limitations of electronic database studies of prescription non-aspirin nonsteroidal anti-inflammatory drugs (NANSAIDs) and risk of myocardial infarction (MI). Pharmacoepidemiol Drug Saf 2005; 14: 513-22

[59] Ulcickas Yood M, Watkins E, Wells KE, et al. Using prescription claims data for drugs available over-the-counter (OTC). Pharmacoepidemiol Drug Saf 2000; 9: S37

[60] Suissa S. Immeasurable time bias in observational studies of drug effects on mortality. Am J Epidemiol 2008; 168: 329-35

[61] McMahon AD. Observation and experiment with the efficacy of drugs: a warning example from a cohort of nonsteroidal anti-inflammatory and ulcer-healing drug users. Am J Epidemiol 2001; 154: 557-62

[62] Greenland S. The effect of misclassification in matched-pair case-control studies. Am J Epidemiol 1982; 116: 402-6

[63] Marshall RJ. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. J Clin Epidemiol 1990; 43: 941-7

[64] Brenner H, Gefeller O. Use of positive predictive value

to correct for disease misclassification in epidemiologic studies. Am J Epidemiol 1993; 138: 1007-15

[65] Motheral B, Brooks J, Clark MA, et al. A Checklist for Retrospective Database Studies-Report of the ISPOR Task Force on Retrospective Databases. Value in Health 2003; 6: 90-7

[66] Tamim H, Tahami Monfared AA, LeLorier J. Application of lag-time into exposure definitions to control for protopathic bias. Pharmacoepidemiol Drug Saf 2007; 16: 250-8

[67] Sjolander A, Humphreys K, Palmgren J. On informative detection bias in screening studies. Stat Med 2008; 27: 2635-50

[68] Ulcickas Yood M, Campbell UB, Rothman KJ, et al. Using prescription claims data for drugs available over-the-counter (OTC). Pharmacoepidemiol Drug Saf 2007; 16: 961-8

[69] Thürigen D, Spiegelman D, Blettner M, et al. Measurement error correction using validation data: a review of methods and their applicability in case-control studies. Statistical Methods in Medical Research 2000; 9: 447-74

[70] Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989; 8: 1051-69

[71] Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within person measurement error. Am I Epidemiol 1992;136: 1400-13

[72] Willett WC, Hunter DI, Stampfer MI, et al. Dietary fat and fiber in relation to risk of breast cancer: an eight year follow-up. JAMA l992 ; 268: 2037-44

[73] Beasley JM, LaCroix AZ, Neuhouser ML, et al. Protein intake and incident frailty in the Women's Health Initiative observational study. J Am Geriatr Soc 2010; 58: 1063-71

[74] Van Roosbroeck S, Li R, Hoek G, et al. Traffic-related outdoor air pollution and respiratory symptoms in children: the impact of adjustment for exposure measurement error. Epidemiology 2008; 19: 409-16

[75] Strand M, Vedal S, Rodes C, et al. Estimating effects of ambient PM(2.5) exposure on health using PM(2.5) component measurements and regression calibration. J Expo Sci Environ Epidemiol 2006; 16: 30-8

[76] Newcombe HB, Kennedy JM, Axford SJ, et al. Automatic linkage of vital records. Science 1959; 130: 954-9

[77] Newcombe H B. Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business. Oxford:Oxford University Press, 1988

[78] Newcombe HB, Fair ME, Lalonde P. The use of names for linking personal records (with discussion). J Am Stat Assoc 1992; 87: 1193-208

[79] Baldi I, Ponti A, Zanetti R, et al. The impact of record-linkage bias in the Cox model. Journal of Evaluation of Clinical Practice 2008; doi:10.1111/j.1365-2753.2009.01119.x

[80] Bohensky MA, Jolley D, Sundararajan V, et al. Data Linkage: A powerful research tool with potential problems. BMC Health Services Research 2010; 10: 346. Available at http://www.biomedcentral.com/1472-6963/10/346

[81] Tromp M, Ravelli AC, Bonsel GJ, et al. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol 2011; 64: 565-72

[82] Silveira DP, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. Rev Saude Publica 2009; 43: 875-82

[83] Grayson DA. Confounding confounding. Am J Epidemiol 1987; 126: 546-53

[84] Weinberg CR. Towards a clearer definition of confounding. Am J Epidemiol 1993; 137: 1-8

[85] Walker AM. Confounding by indication. Epidemiology 1996; 7: 335-6

[86] Bosco JL, Silliman RA, Thwin SS, et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. J Clin Epidemiol 2010; 63: 64-74

[87] Bruce M, Psaty BM, Siscovick DS. Minimizing Bias Due to Confounding by Indication in Comparative Effectiveness Research The Importance of Restriction. JAMA 2010; 304: 897-8

[88] Sturmer T, Schneeweiss S, Avorn J, et al. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol 2005; 162: 279-89

[89] Jackson LA, Jackson ML, Nelson JC, et al. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. Int J Epidemiol 2006; 35: 337-44

[90] Jackson LA, Nelson JC, Benson P, et al. Functional status is a confounder of the association of influenza vaccine and risk of all cause mortality in seniors. Int J Epidemiol 2006; 35: 345-52

[91] Lett HS, Blumenthal JA, Babyak MA, et al. Depression as a risk factor for coronary artery disease: evidence, mechanisms, and treatment. Pychosom Med 2004; 66: 305-15

[92] DiMatteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment. Meta-analysis of the effects of anxiety and depression on patient adherence. Arch Intern Med 2000; 160: 2101-7

[93] White HD. Adherence and outcomes: it's more than taking the pills. Lancet 2005; 366: 1989-91

[94] Brookhart MA, Patrick AR, Dormuth C, et al. Adherence to lipid lowering therapy and the use of preventive health services: an investigation of the healthy user effect. Am J Epidemiol 2007; 166: 348-54

[95] Simpson SH, Eurich DT, Majumdar SR, et al. A meta-

analysis of the association between adherence to drug therapy and mortality. BMJ 2006; 333: 15

[96] Lewis JD, Brensinger C. Agreement between GPRD smoking data: a survey of general practitioners and a population-based survey. Pharmacoepidemiol Drug Saf 2004; 13: 437-41

[97] Dormuth CR, Patrick AR, Shrank WH, et al. Statin adherence and risk of accidents. A cautionary tale. Circulation 2009; 119: 2051-7

[98] Poses RM, Smith WR, McClish DK, et al. Controlling for confounding by indication for treatment: are administrative data equivalent to clinical data? Med Care 1995; 33(Suppl): AS36-46

[99] Chassin MR, Kosecoff J, Park RE, et al. Does inappropriate use explain geographic variation in the use of health care services: a study of three procedures. JAMA 1987; 258: 2533-7

[100] Maynard C, Fisher L, Alderman EL, et al. Institutional differences in therapeutic decision making in the coronary artery surgery study (CASS). Med Decis Making 1986; 6: 127-35

[101] Perrin JM, Homer CJ, Berwick DM, et al. Variations in the rates of hospitalization of children in three urban communities. N Engl J Med 1989; 320: 1183-7

[102] Wennberg JE, Mulley AG, Hanley D, et al. An assessment of prostatectomy for benign urinary tract obstruction: geographic variations and the evaluations of medical care outcomes. JAMA 1988; 259: 3027-30

[103] Goyert GL, Bottoms SF, Treadwell MC, et al. The physician factor in cesarean birth rates. N Engl J Med 1989; 320: 706-9

[104] Wennberg JE. Small area analysis and the medical care outcome problem. In: Sechrest L, Perrin E, Bunker J, eds. Conference proceedings: research methodology-strengthening causal interpretations of non-experimental data. Washington, DC: U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, May 1990

[105] Wennberg JE. Population illness rates do not ex- plain population hospitalization rates. Med Care 1987; 25: 354-9

[106] Glynn RJ, Knight EL, Levin R, Avorn J. Paradoxical relations of drug treatment with mortality in older persons. Epidemiology 2001; 12: 682-9

[107] Schneeweiss S, Wang P. Association between SSRI use and hip fractures and the effect of residual confounding bias in claims database studies. J Clin Psychopharm 2004; 13: 695-702

[108] Roos LL, Fisher ES, Sharp SM, et al. Postsurgical mortality in Manitoba and New England. JAMA 1990; 263: 2453-8

[109] Byar DP. Problems with using observational data-bases to compare treatments. Stat Med 1991; 10: 663-6

[110] Dambrosia JM, Ellenberg JH. Statistical considerations for a medical data base. Biometrics 1980; 36: 323-32

[111] Feinstein AR. Para-analysis, faute de mieux, and the perils of riding on a data barge. J Clin Epidemiol 1989; 42: 929-35

[112] Rothman KJ, Greenland S. Modern epidemiology. 2nd edition. Philadelphia: Lippincott Williams & Wilkins; 1998

[113] Perrio M, Waller PC, Shakir SAW. An analysis of the exclusion criteria used in observational pharmacoepidemiological studies. Pharmacoepidemiol Drug Saf 2006; 16: 329-36

[114] Schneeweiss S, Patrick AR, Sturmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. Med Care 2007; 45(Suppl): S131-42

[115] Petri H, Urquhart J. Channeling bias in the interpretation of drug effects. Stat Med 1991; 10: 577-81

[116] Glynn RJ, Schneeweiss S, Wang P, et al. Selective prescribing can lead to over-estimation of the benefits of lipid lowering drugs. J Clin Epidemiol 2006; 59: 819-28

[117] Danaei G, Tavakkoli M, Hernan MA. Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins. Am J Epidemiol 2012; 175: 250-62

[118] Corrao G, Zambon A, Parodi A, et al. Incidence of cardiovascular events in Italian patients with early discontinuations of antihypertensive, lipid-lowering, and antidiabetic treatments. Am J Hypertens 2012; 25: 549-55

[119] Corrao G, Zambon A, Parodi A, et al. Discontinuation of and changes in drug therapy for hypertension among newly-treated patients: a population-based study in Italy. J Hypertens 2008; 26: 819-24

[120] Corrao G, Conti V, Merlino L, et al. Results of a retrospective database analysis of adherence to statin therapy and risk of nonfatal ischemic heart disease in daily clinical practice in Italy. Clin Ther 2010; 32: 300-10

[121] Corrao G, Romio SA, Zambon A, et al. Multiple outcomes associated with the use of metformin and sulphonylureas in type 2 diabetes: a population-based cohort study in Italy. Eur J Clin Pharmacol 2011; 67: 289-99

[122] Ludvigsson JF, Montgomery SM, Olen O, et al. Coeliac disease and risk of renal disease - a general population cohort study. Nephrol Dial Transplant 2006; 21: 1809-15

[123] Greenland S, Morgenstern H. Matching and efficiency in cohort studies. Am J Epidemiol 1990; 131: 151-9

[124] Klein-Geltink JE, Rochon PA, Dyer S, et al. Readers should systematically assess methods used to identify, measure and analyze confounding in observational cohort studies. J Clin Epidemiol 2007; 60: 766-72

[125] Miettinen OS. Matching and design efficiency in

retrospective studies. Am J Epidemiol 1970; 91: 111-8

[126] Kalish LA. Matching on a non-risk factor in the design of case-control studies does not always result in an efficiency loss. Am J Epidemiol 1986; 123: 551-4

[127] de Graaf MA, Jager KJ, Zoccali C, et. Matching, an appealing method to avoid confounding? Nephron Clin Pract 2011; 118: c315-8

[128] Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 1984; 13: 356-65

[129] Gissler M, Hemminki E. The danger of overmatching in studies of the perinatal mortality and birthweight of infants born after assisted conception. Eur J Obstet Gyn Repr Biol 1996; 69: 73-5

[130] Nordmann S, Biard L, Ravaud P, et al. Case-only designs in pharmacoepidemiology: a systematic review. Plos One 2012; 7: e49444

[131] Maclure M, Fireman B, Nelson JC, et al. W should case-only designs be used for safety monitoring of medical products? Pharmacoepidemiol Drug Saf 2012; 21: 50-61

[132] Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol 1991; 133: 144-53

[133] Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. Biometrics 1995; 51: 228-35

[134] Suissa S. The case-time-control design. Epidemiology 1995; 6: 248-53

[135] Farrington CP. Estimation of vaccine effectiveness using the screening method. Int J Epidemiol 1993; 22: 742-6

[136] Petri H, de Vet HC, Naus J, et al. Prescription sequence analysis: a new and fast method for assessing certain adverse reactions of prescription drugs in large populations. Stat Med 1988; 7: 1171-5

[137] Pariente A, Fourrier-Réglat A, Ducruet T, et al. Antipsychotic use and myocardial infarction in older patients with treated dementia. Arch Int Med 2012; 172: 648-53

[138] Whitaker HJ, Farrington CP, Spiessens B, et al. Tutorial in biostatistics: the self-controlled case series method. Stat Med 2006; 25: 1768-97

[139] Grosso A, Douglas I, Hingorani A, et al. Post-marketing assessment of the safety of strontium ranelate; a novel case-only approach to the early detection of adverse drug reactions. Br J Clin Pharmacol 2008; 66: 689-94

[140] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999; 10: 37-48

[141] Rothwell PM. Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. Lancet 2005; 365: 176-86

[142] Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer-Verlag; 1997

[143] Hosmer DW, Lemeshow S, May S. Applied Survival Analysis. 2nd edition. New Jersey: Wiley; 2008

[144] Zou G. A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol 2004; 159: 702-6

[145] Lumley T, Kronmal R, Ma Shuangge. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. UW Biostatistics Working Paper Series. University of Washington. Paper 293; 2006

[146] Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13-22

[147] Snijders PJ, Boskers R. Statistical treatment of clustered data. Multilevel analysis: an introduction to basic and advanced multilevel modeling. Thousand Oaks, CA: Sage Publications, Inc.; 1999: 13-37

[148] Austin PC, Goel V, van Walraven C. An introduction to multilevel regression models. Can J Public Health 2001; 92: 150-4

[149] Raudenbush SW, Bryk AS. Hierarchical linear models: applications and data analysis methods. Thousand Oaks, CA: Sage Publications, Inc.; 2002

[150] Goldstein H. Multilevel statistical models. 2nd ed. London, UK: Edward Arnold; 1995

[151] Austin PC, Tu JV, Alter DA. Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: should we be analyzing cardiovascular outcomes data differently? Am Heart J 2003; 145: 27-35

[152] Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. Analysis of Longitudinal Data. 2nd edition. New York: Oxford University Press; 2002

[153] Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. New Jersey: Wiley; 2004

[154] Bradbury BD, Gilbertson DT, Brookhart MA, et al. Confounding and control of confounding in nonexperimental studies of medications in patients with CKD. Adv Chron Kidney Dis 2012; 19: 19-26

[155] Seeger JD, Kurth T, Walker AM. Use of propensity score technique to account for exposure-related covariates: an example and lesson. Med Care 2007; 45: S143-8

[156] Takahashi Y, Nishida Y, Asai S. Utilization of health care databases for pharmacoepidemiology. Eur J Clin Pharmacol 2012; 68: 123-9

[157] Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41-55

[158] D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998; 17: 2265-81

[159] Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997; 127: 757-63

[160] Weitzen S, Lapane KL, Toledano AY, et al. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf 2004; 13: 841-53

[161] Shah BR, Laupacis A, Hux JE, et al. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol 2005; 58: 550-9

[162] McWilliams JM, Meara E, Zaslavsky AM, et al. Use of health services by previously uninsured Medicare beneficiaries. N Engl J Med 2007; 357: 143-53

[163] Fu AZ, Liu GG, Christensen DB, et al. Effect of second-generation antidepressants on mania- and depression-related visits in adults with bipolar disorder: a retrospective study. Value Health 2007; 10: 128-36

[164] Schneeweiss S, Maclure M. Use of comorbidity scores for control of confounding in studies using administrative databases. Int J Epidemiol 2000; 29: 891-8

[165] Roos LL, Sharp SM, Cohen MM, et al. Risk adjustment in claims-based research: the search for efficient approaches. J Clin Epidemiol 1989; 42: 1193-206

[166] Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J Clin Epidemiol 1993; 46: 1075-9

[167] Roos LL, Stranc L, James RC, et al. Complications, comorbidities, and mortality: improving classification and prediction. Health Serv Res 1997; 32: 229-38

[168] Von Korff M, Wagner EH, Saunders K. A chronic disease score from automated pharmacy data. J Clin Epidemiol 1992; 45: 197-203

[169] Clark DO, Von Korff M, Saunders K, et al. A chronic disease score with empirically derived weights. Med Care 1995; 33: 783-95

[170] Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992; 45: 613-9

[171] D'Hoore W, Sicotte C, Tilquin C. Risk adjustment in outcome assessment: the Charlson comorbidity index. Methods Inf Med 1993; 32: 382-7

[172] D'Hoore W, Bouckaert A, Tilquin C. Practical considerations on the use of the Charlson comorbidity index with administrative data bases. J Clin Epidemiol 1996; 49: 1429-33

[173] Ghali WA, Hall RE, Rosen AK, et al. Searching for an improved clinical comorbidity index for use with ICD-9-CM administrative data. J Clin Epidemiol 1996; 49: 273-8

[174] Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf 2006; 15: 291-303

[175] Schneeweiss S, Glynn RJ, Tsai EH, et al. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: the example of COX2 inhibitors and myocardial infarction. Epidemiology 2005; 16: 17-24

[176] Phillips CV. Quantifying and reporting uncertainty from systematic errors. Epidemiology 2003; 14: 459-66

[177] Walker AM. Observation an inference. Epidemiology Resources Inc.: Newton, 1991; 120-4

[178] Psaty BM, et al. Assessment and control for confounding by indication in observational studies. J Am Geriatr Soc 1999; 47: 749-54

[179] Kriebel D, Zeka A, Eisen EA, et al. Quantitative evaluation of the effects of uncontrolled confounding by alcohol and tobacco in occupational cancer studies. Int J Epidemiol 2004; 33: 1040-5

[180] Corrao G, Parodi A, Nicotra F, et al. Cardiovascular protection by initial and subsequent combination of antihypertensive drugs in daily life practice. Hypertension 2011; 58: 566-72

[181] Corrao G, Nicotra F, Parodi A, et al. External adjustment for unmeasured confounders improved drug-outcome association estimates based on health care utilization data. J Clin Epidemiol 2012; 65: 1190-9

[182] Steenland K, Greenland S. Monte Carlo Sensitivity Analysis and Bayesian Analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. Am J Epidemiol 2004; 160: 384-90

[183] Sturmer T, Schneeweiss S, Avorn J, et al. Correcting effect estimates for unmeasured confounding in cohort studies with validation studies using propensity score calibration. Am J Epidemiol 2005; 162: 279-89

[184] Sturmer T, Schneeweiss S, Glynn RJ. Performance of propensity score calibration (PSC). Am J Epidemiol 2005; 161(suppl): S75

[185] Bowden, RJ.; Turkington, DA. Instrumental Variables. Cambridge University Press; Cambridge, UK: 1984

[186] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Soc 1996; 91: 444-55

[187] Angrist JD, Krueger AB. Instrumental variables and the search for identification: from supply and demand to natural experiments. J Econ Perspect 2001; 15: 69-85

[188] Murray MP. Avoiding invalid instruments and coping with weak instruments. J Econ Perspect 2006; 20: 111-32

[189] Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects in claims databases using physician-specific prescribing preferences as an instrumental variable. Epidemiology 2006; 17: 268-75

[190] Schneeweiss S, Solomon DH, Wang PS, et al. Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective COX-2

inhibitors and non-selective NSAIDs: an instrumental variable analysis. Arthritis Rheum 2006; 54: 3390-8

[191] Greene WH. Econometric Analysis. 3rd edn. Prentice Hall; Upper Saddle River, NJ: 1997: 740-2

[192] MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the Mediation, Confounding and Suppression Effect. Prev Sci 2000; 1: 173-85

[193] Susser, M. Causal thinking in the health sciences: Concepts and strategies of epidemiology. Oxford University Press; New York: 1973

[194] Breslow NE, Day NE. Statistical methods in cancer research. Volume I—The Analysis of Case-Control Studies. International Agency for Research on Cancer; Lyon: 1980. IARC Scientific Publications No. 32

[195] Meinert CL. Clinical trials: Design, conduct, and analysis. Oxford University Press; New York: 1986

[196] Robins JM. The control of confounding by intermediate variables. Stat Med 1989; 8: 679-701

[197] Last JM. A dictionary of epidemiology. Oxford Unversity Press; New York: 1988

[198] Merchant AT, Pitiphat W. Directed acyclic graphs (DAGs): an aid to assess confounding in dental research. Community Dent Oral Epidemiol 2002; 30: 399-404

[199] Weinberg CR. Toward a clearer definition of confounding. Am J Epidemiol 1993; 137: 1-8

[200] Hernán MA, Hernández-Díaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002; 155: 176-84

[201] Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004; 15: 615-25

[202] Pearl J. Causal diagrams for empirical research. Biometrika 1995; 82: 669-710

[203] Hernández-Díaz S, Schisterman EF, Hernán MA. The birth weight "paradox" uncovered? Am J Epidemiol 2006; 164: 1115-20

[204] James LR, Brett JM. Mediators, moderators and tests for mediation. J Appl Psychol 1984; 69: 307-21

[205] Alwin DF, Hauser RM. The decomposition of effects in path analysis. Am Soc Rev 1975; 40: 37-47

[206] Holland PW. Causal inference, path analysis, and recursive structural equations models. Soc Methodol 1988; 18: 449-84

[207] Sobel ME. Effect analysis and causation in linear structural equation models. Psychometrika 1990; 55: 495-515

[208] Roumie CL, Liu X, Choma NN, et al. Initiation of sulfonylureas versus metformin is associated with higher blood pressure at one year. Pharmacoepidemiol Drug Saf 2012; 21: 515-23

[209] Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 2009; 20: 488-95

[210] Hafeman DM, Schwartz S. Opening the black box: a motivation for the assessment of mediation. Int J Epidemiol 2009; 38: 838-45

[211] Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 1986; 51: 1173-82

[212] Ditlevsen S, Christensen U, Lynch J, et al. The mediation proportion: a structural equation approach for estimating the proportion of exposure effect on outcome explained by an intermediate variable. Epidemiology 2005; 16: 114-20

[213] MacKinnon DP, Lockwood CM, Hoffman JM, et al. A comparison of methods to test mediation and other intervening variable effects. Psychol Meth 2002; 7: 83-104

[214] Taylor AB, MacKinnon D, Tein JY. Test of the three-path mediated effect. Organiz Res Meth 2008; 11: 241-69

[215] Huang B, Sivaganesan S, Succop P, et al. Statistical assessment of mediational effects for logistic mediational models. Stat Med 2004; 23: 2713-28

[216] Schluchter MD. Flexible approaches to computing mediated effects in generalized linear models: generalized estimating equations and bootstrapping. Multivar Behav Res 2008; 43: 268-88

[217] Li Y, Schneider JA, Bennett DA. Estimation of the mediation effect with a binary mediator. Stat Med 2007; 26: 3398-414

[218] Eskima N, Tabata M, Zhi G. Path analysis with logistic regression models: effect analysis of fully recursive causal systems of categorical variables. J Japan Stat Soc 2001; 31: 1-14

[219] Albert JM, Nelson S. Generalized causal mediation analysis. Biometrics 2011; 67: 1028-38

[220] Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004; 15: 615-25

[221] Greenland S. Quantifying biases in causal models: classical confounding vs. collider-stratification bias. Epidemiology 2003; 14: 300-6

[222] Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol 2010; 39: 417-20

[223] Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ 1986; 292: 746-50

[224] Nathan H, Pawlik TM. Limitations of claims and registry data in surgical oncology research. Ann Surg Oncol 2008; 15: 415-23

[225] Sterne JAC, Smith GD. Sifting the evidence—what's wrong with significance tests? BMJ 2001; 322: 226-31

[226] van Walraven C, Austin P. Administrative database

research has unique characteristics that can risk biased results. J Clin Epidemiol 2011; 64: 1-6

[227] International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practices (GPP). Pharmacoepidemiology and Drug Safety 2007; Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/pds.1471

[228] Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965; 58: 295-300

[229] De Roos AJ, Poole C, Teschke K, et al. An application of hierarchical regression in the investigation of multiple paternal occupational exposures and neuroblastoma in offspring. Am J Ind Med 2001; 39: 477-86

[230] Austin PC, Mamdani MM, Juurlink DN, et al. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. J Clin Epidemiol 2006; 59: 964-9

[231] Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990; 1: 43-6

[232] Bender R, Lange S. Adjusting for multiple testingewhen and how? J Clin Epidemiol 2001; 54: 343-9

[233] Efron B, Tibshirani R. Empirical-Bayes methods and false discovery rates for microarrays. Genet Epidemiol 2002; 23: 70-86

[234] Efron B. Large-scale inference. Empirical-Bayes methods for estimation, testing and prediction. Cambridge University Press, New York, 2010

[235] Westfall PH, Tobias RD, Rom D, et al. Multiple Comparisons and Multiple Tests Using SAS. SAS Institute Inc., Cary, NC, 1999

[236] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B 1995; 85: 289-300

[237] Greenland S, Robins J. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology 1991; 2: 244-51

[238] Greenland S. Methods for epidemiologic analysis of multiple exposure: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. Stat Med 1993; 12: 717-36

[239] Greenland S, Poole C. Empirical-Bayes and semi-Bayes approaches to occupational and environmental hazard surveillance. Arch Environ Health 1994; 48: 9-16

[240] Carlin BP, Louis TA. Bayes and empirical Bayes methods for data analysis. Second Edition, New York, Chapman & Hall/CRC,2000

[241] Aickin M. Bayes without priors. J Clin Epidemiol 2004; 57: 4-13

[242] Steenland K, Bray I, Greenland S, et al. Empirical-Bayes adjustment for multiple results in hypothesis-generating or surveillance studies. Cancer Epid Biom Prev 2000; 9: 895-903

[243] Good I. On estimation of small frequencies in contingency tables. J Royal Stat Soc B 1956; 18: 113-24

[244] Efron B, Morris C. Stein's estimation rule and its competitors: an empirical Bayes approach. J Am Stat Ass 1973; 68: 117-30

[245] Greenland S. Principles of multilevel modeling. Int J Epidemiol 2000; 29: 158-67

[246] Hung RJ, Brennan P, Malaveille C, et al. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. Cancer Epid Biom Prev 2004; 13: 1013-21

[247] Witte JS, Greenland S, Haile RW, et al. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. Epidemiology 1994; 5: 612-21

[248] Bagnardi V, Botteri E, Corrao G. Empirical-Bayes adjustment improved conventional estimates in postmarketing drug-safety studies. J Clin Epidemiol 2006; 59: 1162-8

[249] Petitti DB, Teutsch SM, Barton MB, et al. Update on the methods of the U.S. Preventive Services Task Force: insufficient evidence. Ann Intern Med 2009; 150: 199-205

[250] Doust J. Why do doctors use treatments that do not work? BMJ 2004; 328: 474

[251] Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003; 290: 1624-32

[252] Slutsky JR, Clancy CM. Patient-centered comparative effectiveness research: essential for high-quality care. Arch Intern Med 2010; 170: 403-4

[253] Dreyer NA, Schneeweiss S, McNeil BJ, et al. Research Support, Non-U.S. Gov't United States. Am J Manag Care 2010; 16: 467-71

[254] Inst. Med. (IOM). 2009. Initial National Priorities for Comparative Effectiveness Research.Washington, DC:Natl. Acad. Available at. http://www.nap.edu/catalog/12648.html

[255] Institute of Medicine (IOM) 2011. Clinical Practice Guidelines that We Can Trust. Washington, DC: Natl. Acad. Available at. http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust.aspx

[256] Garrison LP, Neumann PJ, Erickson P, et al. Using real-world data for coverage and payment decisions: The ISPOR Real-World Data Task Force Report. Value in Health 2007; 10: 325-35

[257] Zhang J, Yun H, Wright NC, et al. Potential and pitfalls of using large administrative claims data to study the safety of osteoporosis therapies. Curr Rheumatol Rep 2011; 13: 273-82