

RESEARCH ARTICLE

Open Access



# Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review

Guillermo Prada-Ramallal<sup>1,2</sup>, Bahi Takkouche<sup>1,2,3</sup> and Adolfo Figueiras<sup>1,2,3\*</sup> 

## Abstract

**Background:** The availability of clinical and therapeutic data drawn from medical records and administrative databases has entailed new opportunities for clinical and epidemiologic research. However, these databases present inherent limitations which may render them prone to new biases. We aimed to conduct a structured review of biases specific to observational clinical studies based on secondary databases, and to propose strategies for the mitigation of those biases.

**Methods:** Scoping review of the scientific literature published during the period 2000–2018 through an automated search of MEDLINE, EMBASE and Web of Science, supplemented with manually cross-checking of reference lists. We included opinion essays, methodological reviews, analyses or simulation studies, as well as letters to the editor or retractions, the principal objective of which was to highlight the existence of some type of bias in pharmacoepidemiologic studies using secondary databases.

**Results:** A total of 117 articles were included. An increasing trend in the number of publications concerning the potential limitations of secondary databases was observed over time and across medical research disciplines. Confounding was the most reported category of bias (63.2% of articles), followed by selection and measurement biases (47.0% and 46.2% respectively). Confounding by indication (32.5%), unmeasured/residual confounding (28.2%), outcome misclassification (28.2%) and “immortal time” bias (25.6%) were the subcategories most frequently mentioned.

**Conclusions:** Suboptimal use of secondary databases in pharmacoepidemiologic studies has introduced biases in the studies, which may have led to erroneous conclusions. Methods to mitigate biases are available and must be considered in the design, analysis and interpretation phases of studies using these data sources.

**Keywords:** Pharmacoepidemiology, Observational studies, Bias, Confounding factors, Medical records, Electronic health records, Administrative claims, Medical record linkage

## Background

In recent decades, with advances of computer technology and the exponential growth in the quantity of data available, new opportunities for research in many fields have emerged. One of these fields is the health sector, due to the availability of clinical and therapeutic data drawn from medical records and administrative databases used for

billing and other fiscal functions related to the provision of patient care (i.e. secondary databases) [1].

This availability of data has increased the interest of pharmacoepidemiologists in using secondary databases as sources of data for research. Contributing to this is the perception that clinical trials are not always useful for evaluation of therapies in real-world practice, particularly those providing limited safety data. However, swift and easy access to this information may be deceptively simple [2]. Indeed, the utilization of secondary databases entail not only the limitations specific to observational epidemiologic research but those inherent to these specific types of sources [3], as well as the social and ethical challenges related to data privacy and security [4, 5].

\* Correspondence: [adolfo.figueiras@usc.es](mailto:adolfo.figueiras@usc.es)

<sup>1</sup>Department of Preventive Medicine and Public Health, University of Santiago de Compostela, c/ San Francisco s/n, 15786 Santiago de Compostela, A Coruña, Spain

<sup>2</sup>Health Research Institute of Santiago de Compostela (Instituto de Investigación Sanitaria de Santiago de Compostela - IDIS), Clinical University Hospital of Santiago de Compostela, 15706 Santiago de Compostela, Spain  
Full list of author information is available at the end of the article



Consequently, many researchers recommend caution and warn against the high risk of introducing biases when using these databases [6–9]. The aim of this study was thus to review the literature of the last two decades in which the authors highlight the existence of some type of bias in observational clinical studies based on secondary data sources, in order to identify the most common biases and explore the perception of this issue in the pharmacoepidemiologic field over time and across medical research disciplines. We then propose possible strategies to control the biases identified in the review.

## Methods

We carried out a scoping review, which is a methodological strategy that enables the results of an exploratory research to be summarized. In this type of review, unlike other systematic reviews, the application of quality filters is not an initial priority [10]. We performed and reported our study based on the methodological guidance for the conduct of a scoping review from the Joanna Briggs Institute [11] and the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) Extension guideline for Scoping Reviews [12]. The protocol for this scoping review is available on request from the corresponding author.

### Data-sources and search strategy

An automated search of bibliographic databases was performed, with an initial search in MEDLINE, subsequently supplemented by EMBASE and Web of Science. To avoid duplicated results, in EMBASE and Web of Science we used the option that enables journals indexed in MEDLINE to be excluded. The same free-text search strategy was applied in the 3 databases: (*clinical-data\* OR health-data\* OR medical-data\* OR prescription-data\* OR administrative-data\* OR epidemiologic-data\* OR health-claim\* OR administrative-claim\* OR insurance-claim\* OR claims-data\* OR health-record\* OR medical-record\**) AND (*confounding OR bias\* OR missing-data OR misclassification*) AND (*observational OR epidemiolog\* OR pharmacovigilance OR challenge\**) AND *drug*, from January 1, 2000 to January 1, 2018. All types of research design were considered. Adding restrictive MeSH (Medical Subject Headings) terms according to type of publication was not deemed suitable, since this was found to lead to an excessive reduction in search sensitivity.

Once the references were identified, the titles and the abstracts, when available, were used as a preliminary screening filter, and if deemed potentially relevant, full text articles were retrieved. Other relevant references were identified by manually cross-checking reference lists of selected articles and using the “related articles” option. This full screening was performed by two reviewers (GP-R, AF). Discrepancies were discussed between the two reviewers to

achieve consensus. In case of a possible disagreement, a third author (BT) was designated.

### Article selection and data abstraction

We included in the review opinion essays, methodological reviews, analyses/reanalyses and simulation studies, as well as letters to the editor or retractions, the principal objective of which, described in their abstracts, was to highlight the existence of some type of bias in pharmacoepidemiologic studies that used secondary health care databases.

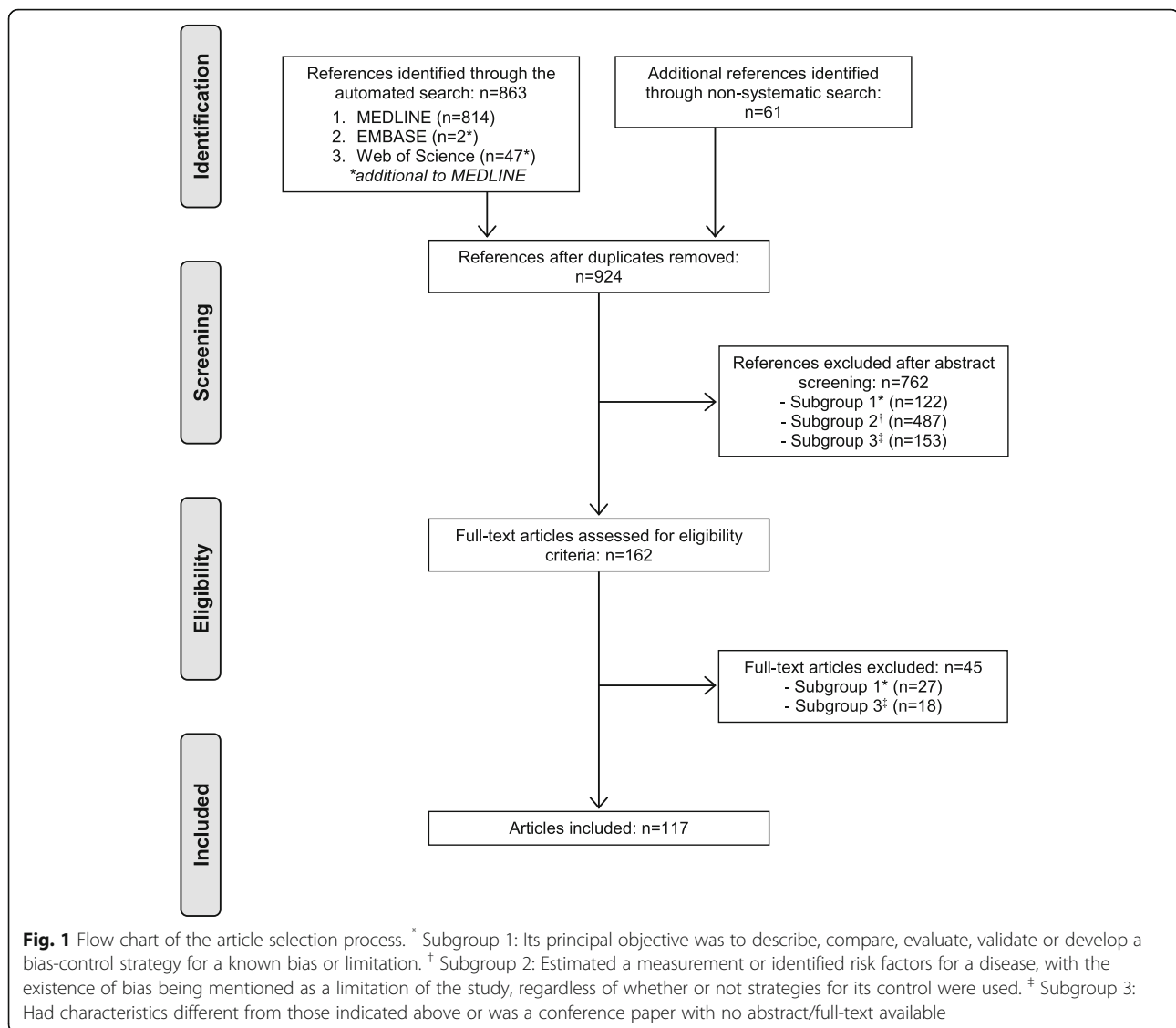
In order to reduce the number of identified references and thus simplify the display of the results, the following exclusion criteria were considered that classified dismissed references into subgroups: (1) its principal objective was to describe, compare, evaluate, validate or develop a bias-control strategy for a known bias or limitation (e.g. analytical method, study design, algorithm, framework); (2) it estimated a measurement (e.g. association treatment-effect) or identified risk factors for a disease, with the existence of bias being mentioned as a limitation of the study, regardless of whether or not strategies for its control were used; (3) it had characteristics different from those indicated above (e.g. studies with different objectives, not based on secondary databases, with no drug involved, no bias mentioned) or it was a conference paper with no abstract/full-text available.

A data charting form was jointly developed by two reviewers (GP-R and AF) to determine which variables had to be extracted. One person (GP-R) extracted the information from the articles (i.e. first author, publication date, category under which the journal was indexed –if the journal was indexed under more than one category, the category under which it was best ranked was considered–, type of article, type of bias(es) mentioned) and when further clarification was needed, articles were checked and validated by additional reviewers as a form of quality control (AF and BT). The three reviewers discussed the results and continuously updated the data charting form.

The synthesis included both quantitative analysis (i.e. publication trend of identified/included articles and frequency analysis of the biases mentioned) and qualitative analysis (i.e. content analysis) of the components of the research purpose.

## Results

Figure 1 shows the article selection process. A total of 117 articles were included. The automated search resulted in the identification of 863 non-duplicated references, which were reduced to 56 after application of the exclusion criteria. The manual selection process incorporated a further 61 references.



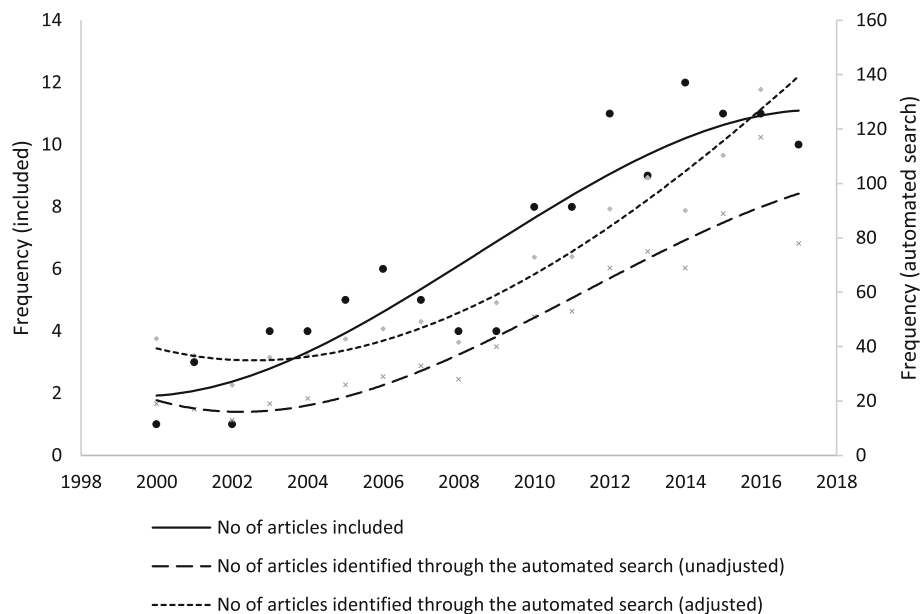
### Publication trend

Figure 2 shows a polynomial smoothing of the frequency with which the articles included in the review were published since 2000. An increasing trend is observed, so that nearly half (45.3%, 53/117) of the articles were published during the last 5 full years of this review. There is a similar trend in the timeline of references identified through the automated search when adjusted by the number of indexed citations added to MEDLINE during each year [13], which suggests that the restriction criteria considered did not introduce any selection bias. A slight decrease in 2017 may be due to inherent characteristics of the indexing process in the bibliographic databases, or to the fact that the most recent references have had less time to be cited, and consequently are less likely to be identified by the cross-reference manual search.

There seems to be a wide variety of disciplines interested in articles about the potential limitations of secondary databases (see Fig. 3a). Overall, the most frequently used categories of medical journals were “Public, environmental & occupational health” (24.8%, 29/117 articles included) and “Pharmacology & pharmacy” (14.5%, 17/117). In general, the same publication trend over time is observed when stratifying by discipline (see Fig. 3b).

### Major biases mentioned in the articles included in the review

Table 1 lists the articles that mentioned the categories or subcategories of the biases most usually described in observational studies of pharmacoepidemiologic databases. Confounding bias as such, or in any of its diverse forms of presentation, was the most frequently mentioned category of bias (63.2%, 74/117 articles included), while confounding



**Fig. 2** Publication timeline of the 117 articles included in the review (left Y axis) and the 863 references identified through the automated search (right Y axis) unadjusted and adjusted by the number of indexed citations added to MEDLINE

by indication was the most frequent subcategory (32.5%, 38/117) followed by unmeasured/residual confounding (28.2%, 33/117). Mention was also made of time-dependent confounding and over-adjustment due to inappropriate choice of variables in the statistical model (bias from misspecification of control variables).

Similarly, some type of selection and measurement bias was mentioned in 47.0% (55/117) and 46.2% (54/117) of the articles included, respectively. Bias due to missing data and prevalent user bias were the most frequently reported selection biases (38.2%, 21/55 and 21.8%, 12/55, respectively); in addition, other forms of bias were also described, such as protopathic bias, informative censoring, competing risks, and differential health care access bias. Exposure or outcome misclassification were the most usual causes of measurement bias (51.9%, 28/54 and 61.1%, 33/54 respectively). Temporal ambiguity and misclassification of confounders were likewise cited.

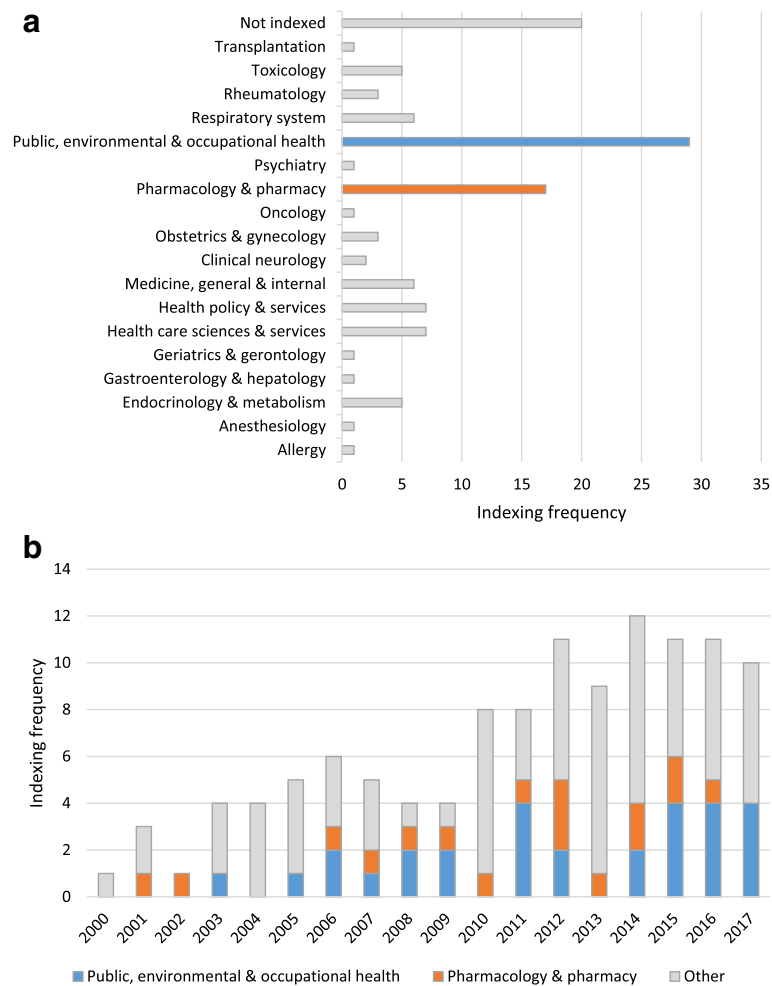
Although they can strictly be considered a subset of the larger 3 categories (i.e. confounding, selection or measurement bias), last to be examined was the category of time-related biases, such as the “immortal time” bias, which proved to be the single most reported bias (25.6%, 30/117) after confounding by indication, unmeasured/residual confounding and outcome misclassification (28.2%, 33/117). Immeasurable time bias, time-window bias and time-lag bias were also described. Figure 4 shows the frequency for each bias mentioned in the articles included, as well as the overarching categories, stratified by 6-year time periods.

Additional file 1: Table S1 contains the data extracted from the included articles in descending order of publication date by the research field category under which the journal was indexed. The articles were also classified according to type of content, including, in each case, the categories or subcategories of bias mentioned.

## Discussion

This is the first known structured review that explores potential biases in observational studies of pharmacoepidemiologic databases. The results of this review suggest that there is growing concern in the scientific literature about identifying, describing and controlling such biases. This should not be overlooked, since observational epidemiologic database studies currently afford an excellent opportunity for medical research. The results of these studies are to be valid and applicable to decision-making about safety and effectiveness. It is then of paramount importance that proper account be taken of these biases to ensure that they are correctly controlled for.

Confounding bias as such, or in any of its diverse forms of presentation, is mentioned in almost two-thirds of the articles included in the scoping review (see Table 1 for references). Adequate control of confounding poses a challenge in studies that use health care databases, since these were not designed for undertaking epidemiologic studies. The absence or poor quality of data on potential confounding factors in secondary databases (e.g. over-the-counter drugs, frailty of the subject, smoking habit) is a frequent phenomenon [14–17], which renders



**Fig. 3** a Distribution of included articles across medical disciplines. b Timeline of included articles by most prevalent indexed disciplines

it difficult or even impossible to adjust for such factors, in order to control for confounding [18].

If data on confounding variables has been collected, the reviewed articles propose different control methods: (1) in the design stage, through the application of restriction criteria, matching methods, or implementation of a new-user design (see below, *depletion of susceptibles*); and (2) in the analysis stage, through stratification of patients across treatment groups according to relevant factors, or multivariate regression techniques, by including these confounding factors as independent variables in regression models. In cases in which the number of variables is very high, adjusting for the disease risk score [19] or the propensity score to receive treatment may be of interest [20, 21].

Among the studies dealing with the issue of confounding in pharmacoepidemiology, the most commonly described type of confounding is *confounding by indication for treatment* (the treatment decision is associated with an indication, which is in turn a risk factor for the disease), which is mentioned in one-third of the articles

reviewed (see Table 1). Confounding by indication, often also referred to as channeling bias, is closely related to selection bias [22]. Some useful analytical control methods proposed include separating the effects of a drug taken at different times [23], sensitivity analysis for unmeasured confounding factors (see below), and the use of instrumental variables [24]. Furthermore, according to the literature reviewed, there seems to be a general agreement that conventional methods for control of confounding factors are inadequate in controlling *time-dependent confounding* (mentioned in 6.0% of the articles reviewed, see Table 1). G-estimation [25] and marginal structural models [26] are alternative methods for achieving such control.

More than a quarter of the articles included in the scoping review consider the absence of quality data to control for potential confounding variables as an important limitation of observational pharmacoepidemiologic studies using secondary databases (see Table 1). Therefore, the proposed strategies for the control of *unmeasured variables* include

**Table 1** Articles that mention the most usual biases described in observational studies of pharmacoepidemiologic databases

Category/Subcategory	Description of the bias	References (n = 117)	Percentage (%)
Confounding	The measure of association between treatment and outcome is distorted by the effect of one or more variables, which are also risk factors for the outcome of interest	[1–3, 6, 14–16, 18, 22, 40, 41, 57, 58, 62, 80–139]	63.2
Confounding by indication <sup>a</sup>	The clinical condition that determined the prescription of the treatment is associated with the effect, acting as a confounding factor (e.g. a worse disease status at baseline: confounding by disease severity)	[3, 6, 18, 22, 40, 41, 57, 80, 82, 84, 86, 87, 89, 90, 92, 96, 97, 99, 100, 104, 106, 107, 110, 111, 113, 114, 116, 118, 120, 122, 126, 128–131, 133, 134, 138]	32.5
Time-dependent confounding	A variable that can vary with time acts as a confounding factor between the current exposure and outcome, and as an intermediary between prior and current exposure	[40, 41, 57, 58, 81, 92, 104]	6.0
Unmeasured/residual confounding	There is not enough information about all the relevant confounding factors known, unknown or difficult to measure (e.g. frailty). If confounding cannot be completely controlled for, the residual confounding effect of some factors remains in the final effect that is observed	[1–3, 6, 14, 15, 18, 58, 62, 80–83, 86, 89, 91–93, 96, 101, 103, 108, 110, 113, 116, 119, 125, 127, 130, 132, 134, 136, 139]	28.2
Healthy user/adherer effect	Access to health care resources is associated with a higher level of education and health-seeking behavior. Furthermore, patients who comply with the treatment during prolonged periods of time tend to be healthier	[2, 18, 91, 96, 125, 127]	5.1
Selection bias	The study sample population is not representative of the target population to which the results will be extrapolated	[2, 16, 18, 22, 40, 41, 54, 57, 58, 63, 81, 83, 84, 87, 88, 90, 91, 93–95, 99, 101–103, 105, 107–109, 111–113, 115–119, 121, 122, 124, 125, 135–137, 140–151]	47.0
Protopathic bias	The treatment is associated with subclinical disease stages (an early manifestation of the still undiagnosed condition under study gives rise to prescription of the treatment)	[40, 41, 81, 109]	3.4
Losses to follow-up (informative censoring)	The mechanism that triggers discontinuity of the treatment is associated with the risk of observing the outcome of interest	[40, 41, 116]	2.6
Depletion of susceptibles (prevalent user bias)	The inclusion of prevalent instead of incident users entails insufficient verification of the adverse effects that occur at the beginning of treatment (those susceptible to the adverse effect have interrupted the treatment)	[2, 40, 41, 57, 83, 90, 99, 107, 111, 116, 118, 148]	10.3
Missing data	In multivariate analyses, such as regression models, observations that lack one or more of the values of a variable included in the model tend to be eliminated	[58, 63, 87, 93, 94, 108, 112, 116, 119, 125, 135–137, 140, 141, 143–147, 151]	17.9
Measurement bias	Data on true exposures, outcomes and other variables are recorded in the form of indicators (observed measures) that do not accurately reflect reality	[2, 3, 6, 7, 16, 40, 41, 54, 55, 58, 87, 88, 91, 93, 94, 96, 101, 105, 108, 110, 112, 114, 115, 117, 119, 121, 124, 125, 130, 135–138, 140, 141, 143, 144, 146, 147, 149, 151–164]	46.2
Misclassification bias	The association between treatment and outcome is distorted by systematic errors, due to the way in which the variables of interest are measured in comparison groups	[2, 3, 6, 7, 16, 40, 41, 54, 55, 58, 87, 88, 91, 93, 94, 96, 101, 105, 108, 110, 112, 114, 115, 119, 121, 125, 130, 135–138, 140, 141, 143, 144, 146, 147, 149, 152–164]	43.6
Misclassification of exposure	The measure of exposure of a given treatment is not an exact reflection of its real use (e.g. flawed measurement, non-compliance with treatment, inappropriate use of time windows)	[2, 3, 16, 40, 41, 54, 55, 58, 87, 91, 93, 94, 96, 101, 110, 119, 121, 130, 138, 140, 146, 147, 152, 154, 156, 158, 159, 164]	23.9
Misclassification of outcome	Error in the diagnosis (e.g. clinical ambiguity, non-uniform coding)	[2, 3, 6, 7, 16, 40, 41, 54, 58, 87, 91, 93, 94, 96, 101, 110, 112, 114, 121, 125, 135–137, 141, 143, 149, 153, 155, 157, 160–163]	28.2



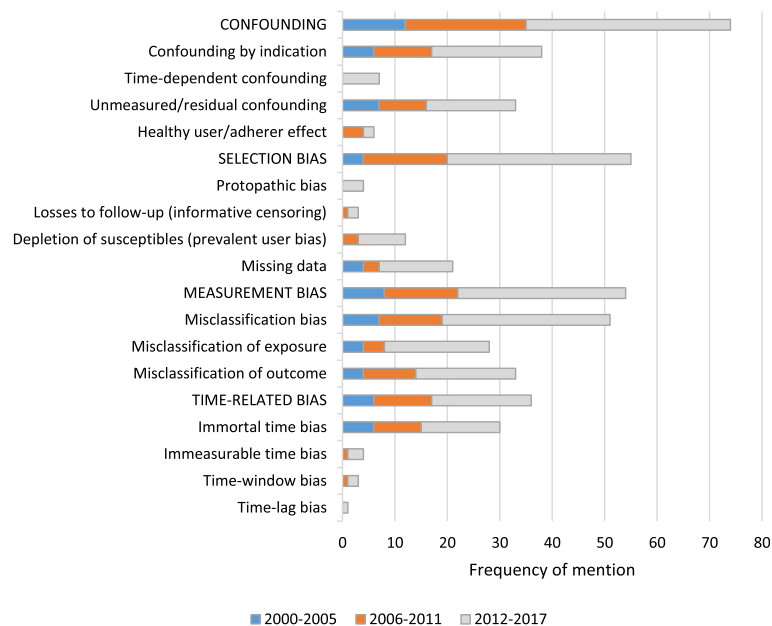
**Table 1** Articles that mention the most usual biases described in observational studies of pharmacoepidemiologic databases (Continued)

Category/Subcategory	Description of the bias	References (n = 117)	Percentage (%)
Time-related bias	Follow-up time and exposure status are inadequately taken into account in the study-design or analysis stages	[2, 7, 40, 41, 57, 68–75, 77, 83, 86, 87, 90, 99, 101, 105–107, 111, 114, 118, 128, 129, 133, 142, 165–170]	30.8
Immortal time bias	A period of time (immortal) during which the study event cannot occur is included in the follow-up or is excluded from analysis due to an incorrect definition of the start of follow-up	[2, 7, 40, 41, 57, 68–75, 77, 83, 86, 87, 90, 99, 101, 106, 107, 111, 114, 118, 128, 129, 133, 166, 167]	25.6
Immeasurable time bias	A period of time (immeasurable) during follow-up is ignored and thus misclassified as unexposed period, since outpatient prescriptions that define exposure cannot occur (e.g. serious chronic diseases that require extensive use of medications and multiple hospitalizations)	[142, 165, 168, 170]	3.4
Time-window bias	The use of time-windows of different lengths between cases and controls to define time-dependent exposures prevents subjects from having the same opportunity time to receive prescriptions	[90, 106, 169]	2.6
Time-lag bias	Comparisons are conducted of treatments given at different stages of the disease, which inherently introduces bias related to disease duration and progression	[106]	0.9

<sup>a</sup>Sometimes also referred to as channeling bias

the performance of sensitivity analyses and use of information external to the database [27–29]. Instrumental variable techniques, proxy measures and propensity scores, excluding from the analysis treated and untreated subjects having extreme values, have also been used [30]. In the design stage, case-crossover study designs, where each study participant receives all treatments that are being investigated

but at different times [31], and restriction to an active comparison group can be useful. The active comparator design emulates the design of a head to head randomized controlled trial. Instead of using a non-user group, the drug of interest is compared with another drug commonly used for the same indication. By ensuring that treatment groups have similar characteristics, this design potentially helps to

**Fig. 4** Frequency of the biases mentioned in the included articles stratified by time periods

mitigate both measured and unmeasured confounding [32]. At all events, with the exception of crossover designs, where the order in which a study participant receives the treatments is randomized, control for unmeasured variables will never be optimal or, at best, one could never be sure that it would be so. But even in this case, the crossover design may still be affected by time-dependent confounding.

In this context, Hernán has proposed a new approach based on the use of observational data from a large health care database to emulate a hypothetical randomized trial (the target trial) [33]. Although the emulated target trial helps avoid common methodologic pitfalls, the appropriate adjustment for time-dependent confounders remains critical [34].

In contrast to clinical trials, an advantage of observational pharmacoepidemiologic studies in which the study populations are constructed on the basis of large health care databases is the inclusion of *frail patients*. However, some authors have argued that due to the fact that frailty is difficult to measure and a strong risk factor for unfavorable outcomes, it will lead to unmeasured and residual confounding, and possibly to paradoxical results [35, 36]. Frailty is an example of an unmeasured confounding variable [14, 15].

About 5% of the reviewed articles deal with the *healthy user effect* (see Table 1), which consists of a type of confounding generated because patients with healthier behaviors generally demand medical attention more frequently for preventive treatments or asymptomatic chronic diseases. These patients are also more likely to be better adherers. Accordingly, part of the apparent efficacy/safety of the treatment will be due, not to the treatment per se, but rather to the healthier behaviors that are associated with those taking it [18, 37]. In observational studies of pharmacoepidemiologic databases, these types of behavior are seldom measured, thus making it very difficult to control for their effect [38].

Almost half of the articles included in the scoping review mention some type of selection bias. Within this category, it is worth highlighting the *protopathic bias*. Although this bias is not widely mentioned in our review (3.4%, see Table 1), possibly because it is unusual for the treatment to be associated with subclinical states and/or early symptoms of the disease, the impact of this bias may be important. However, controlling protopathic bias is not easy since it is not a confounding bias, and adjustment techniques are thus useless. In this case, we must resort to restriction of the exposure group to patients with indications that are unrelated to the initial states of the disease under study. Another option for controlling protopathic bias is to use the concept of lag-time to define the etiologic window in which the exposure to the drug is assessed [39].

Consumption of medicines under real conditions is subject to important variations (e.g. variation in the dose, treatment interruptions, dropouts), especially in the management of chronic diseases. This variability may be due to changes in the disease (increasing or decreasing severity) or in the effect of the drug (adverse events or interactions). The traditional approach through an “as-treated” analysis, in which one censors subjects who interrupt their treatment during follow-up, may introduce bias since censored subjects (losses to follow-up) are systematically at higher or lower risk of developing the outcome [40, 41]. In practice, this *informative censoring* (mentioned in only 2.6% of the articles reviewed, see Table 1) leads to a selection bias. For example, if the clinical effects expected are not met then the treatment is suspended or modified. The bias consists in selecting for the analysis data of patients for whom the treatment produces the expected outcome [42]. This bias may be identified through sensitivity analyses. In this regard, the use of databases represents an important advantage as information on the outcome may be available even when the treatment was suspended. To control the bias introduced by an exposure to the drug that varies with time, it could prove useful to consider that exposure as a time-dependent variable in an appropriate multivariate regression model. Procedures based on the inverse probability censoring weighting have also been proposed [43].

Judging by the number of articles that mention it (10.3%), greater importance has been given to another type of selection bias known as *depletion of susceptibles*, which is caused by the inclusion in the study of both prevalent and incident treatment users (see Table 1). Prevalent users (“survivors” from the first treatment period) may not have the same risk of an adverse event as incident (new) users, i.e., those who tolerate the medication continue using it and those who do not tolerate the medication (susceptible to the adverse event) have stopped using it. This bias can be prevented in the design stage of the study by limiting the follow-up to new users [44]. The new-user design allows potential confounding factors to be measured just before the start of follow-up. This way, these confounding factors will not be affected by the treatment. Adjustment for differences between treatment groups will then use the baseline values of the confounders [45].

Apart from ensuring an appropriate adjustment for confounding, the new-user design potentially reduces immortal time bias (see below) when combined with the active comparator design by implementing similar definitions of the index date across comparison groups [32]. The new user design combined with the active comparator design can also reduce confounding by indication and other unmeasured patient characteristics (e.g. frailty, healthy user) at the design stage [46].

As our results suggest, one of the major challenges in the analysis of observational data is the *missing data*



issue [47], which is mentioned in almost one of every five articles included in the scoping review (see Table 1). If the probability of missing an observation is independent of both observed and missing data, complete cases are assumed to be a random sample of the full dataset (i.e. missing completely at random [48]). In this case, dropping cases with missing data may give unbiased estimates. However, in the multivariate analysis, observations (or subjects) are eliminated whenever where data of a variable included in the model are missing. As a consequence, observations with missing values may lead to a substantial attrition of the sample size. If this lack of information is associated with an important characteristic (e.g. severity, frailty), an effect equivalent to selection bias is produced.

Sometimes, it is assumed that the probability of missing an observation may be predicted by variables that are measured previously, but which are not further dependent on unmeasured variables (i.e. missing at random [48]). That is, the probability of dropout will depend on observed values. Although standard analysis of the available cases is potentially biased in this case, methods that can provide valid analysis are available, but these require additional appropriate statistical modeling.

In both circumstances described above, likelihood-based methods (e.g. mixed models), in which missing data can be estimated using the conditional distribution of the other variables, can be useful for controlling bias [49]. There are alternative techniques, such as multiple imputation, that preserve the natural variability of the data [50] and incorporate the uncertainty due to missing data [51], with which similar results are obtained. Inverse probability weighting (where complete cases are weighted by the inverse of their probability of being a complete case) is also a commonly used method to reduce this bias. While multiple imputation requires a model for the distribution of missing data given the observed data, the inverse probability weighting requires a model for the probability that an individual is a complete case [52]. In any case, it is important that all covariates on which missingness depends be included in the model.

On the contrary, if the fact that an observation is missing is predicted by unmeasured variables, such as the outcome of interest (i.e. missing not at random, sometimes called “non-ignorable non-response” or “informative missingness”), then no statistical approach can give unbiased estimates. When missingness cannot be empirically modelled, the recommended approach is to conduct sensitivity analyses to determine the extent of missingness [53].

After confounding by indication and unmeasured/residual confounding, our results show that the bias most frequently described in studies using secondary health care databases is that due to systematic misclassification errors which distort the association between treatment and outcome. *Exposure or outcome misclassification*, which is

mentioned in almost half of the articles included in the scoping review (see Table 1), can give rise to measurement biases and heterogeneity [17, 54, 55]. To prevent this, a validation study of these variables should first be conducted, followed by the performance of a sensitivity analysis or application of regression techniques [56]. Medical records are normally considered the *gold standard* or reference for intermediate and final outcome variables but display limitations in the recording of all medications taken by patients [57]. While dispensing records are more detailed in measurement of exposure (though they do not record the over-the-counter or out-of-pocket consumption at an individual level), they nonetheless lack outcome variables [1, 3, 58, 59]. It is therefore important to link both types of data sources [60, 61] and consider, when necessary, the use of additional data collected expressly for research purposes [15, 62, 63], to avoid errors that may generate misleading conclusions [64, 65].

The last category of bias identified was that related to time. However, it must be taken into account that the mechanism that underlies the generation of a time-related bias may be closely related to the other larger categories described (i.e. confounding, selection or measurement bias). By far, the most frequently described time-related bias is the *immortal time bias*, which is mentioned in one of every four articles reviewed (see Table 1). Immortal time bias (where the follow-up includes a time period during which the study event cannot occur or is excluded from the analysis due to an incorrect definition of the start of follow-up) resurged with a number of observational studies that reported surprisingly beneficial effects of drugs [66, 67] and is increasingly being described in cohort studies of pharmacoepidemiologic databases [68–70]. Suissa warns about the risk of reporting absurd conclusions, if inappropriate data-analysis methods are used [69–75]. To prevent this, the entire follow-up time, including that preceding the start of exposure, must be considered, and exposure during immortal time must be correctly classified [76]. By applying a Cox model with time-dependent exposures, more reliable estimates can be obtained [69, 77, 78].

### Limitations

This scoping review presents the limitations inherent to this type of study design. In contrast to classical systematic reviews, the aim of which is to provide answers to a clearly defined research question, the scoping studies are less likely to seek very specific research questions nor, consequently, to assess the quality of included studies [79]. In this sense, a potential reviewer's bias in the assessment of the restriction criteria cannot be ruled out since they are not based on a measurable quality of the identified references. However, we do not believe that this may hinder the purpose and the conclusions of the review.

**Table 2** Main bias-control strategies in observational studies of pharmacoepidemiologic databases (Continued)

Category	Control strategies
	with the same indication, or limiting participation to patients without a risk factor for the effect that could have determined the treatment
Time-dependent confounding	- G-estimation - Marginal structural models
Unmeasured confounding	- Crossover design - Asymmetric exclusion of patients with extreme propensity-score values - Instrumental variables - Proxy measures - Restriction (active comparison group) - Sensitivity analysis - Validation study + external adjustment
Selection bias	
Protopathic bias	- Restriction (e.g. restricting the untreated group to a population with the same indication, or restricting the treated group to a population with an indication that is not a subclinical stage of the disease) - Excluding a specific period of time prior to the date of diagnosis of the disease (lag-time) from the etiologic window
Losses to follow-up (informative censoring)	- Inclusion of variables that affect censoring and event times in the multivariate regression model - Inverse probability of censoring weighting - Sensitivity analysis
Depletion of susceptibles (prevalent user bias)	- New-user design - Meta-analysis of clinical trials
Missing data	- Replacing each absent observation with a mean value based on observed values of the variable or the predicted value based on a regression model - Imputation methods (e.g. multiple imputation) - Likelihood-based methods - Inverse probability weighting
Measurement bias	
Misclassification bias	- Validation study (exposure/outcome/confounders) + (sensitivity analysis/misclassification control techniques using multivariate regression)
Time-related bias	
Immortal time bias	- Data analysis with procedures that take into account time-dependent exposure in a cohort - Transferring the start of treatment to the end of the immortal time period in both groups
Immeasurable time bias	- Data analysis accounting for the time-varying exposable period
Time-window bias	- Accounting for duration of treatment in the selection of controls - Time-dependent analysis
Time-lag bias	- Comparing patients at the same stage of disease

**Table 2** Main bias-control strategies in observational studies of pharmacoepidemiologic databases

Category	Control strategies
Confounding	
Measured confounding	- Multivariate analysis - Restriction* - Stratification - Matching - New-user design - Propensity score - Large-scale, simple randomized trials - Meta-analysis of clinical trials * <i>Confounding by indication</i> : Restricting the untreated group to a population

Due to the exploratory nature of this review, its purpose was not to obtain all available evidence on a specific topic, but rather that from a subset of the literature on a broad topic (bias in observational pharmacoepidemiologic studies using secondary data sources), where many different study designs might be applicable (opinion essays, methodological reviews, analyses, letters to the editor or retractions). Although a wide-search strategy was employed, some relevant studies may have been missed. Therefore, the existence of some selection bias cannot be ruled out. Furthermore, the search strategy itself, intentionally designed to identify articles that highlight the limitations of secondary databases, does not allow an unbiased comparison with the articles that may show the advantage of secondary databases.

Given the above limitations, and the fact that information on bias was extracted based on the description provided by the original authors, another limitation would be related to the quantification of each type of bias. This should be interpreted as an approximate measure of the impact of the bias on the published literature (i.e. what is prominently talked about), but not as an estimate of the probability of occurrence (or detection) of the bias in the population of pharmacoepidemiologic studies that use secondary databases, since it may be influenced by the ease of describing that specific bias or by the interest that the bias may have raised in the studies of the most prolific authors in the field (e.g. immortal time bias). It is therefore possible that a certain degree of misclassification of some biases exists.

## Conclusions

The emergence of health care databases has caused dramatic changes in pharmacoepidemiology. Due to routine, automated capture of data on drug prescription and dispensing that are used for administration purposes, together with the implementation of electronic medical records, secondary databases have generated enormous possibilities and expectations about their potential. This happens, moreover, at a time when it is recognized that

clinical trials cannot answer questions about the effectiveness and safety of treatments in clinical practice.

Superficially, secondary databases afford the possibility of performing studies rapidly, at low cost, with enormous sample sizes, objective data and long-term follow-up. Even so, their limitations should not be ignored. This review provides a complete overview of the potential biases inherent to this type of data sources, including the weighting of their impact on the literature of the last two decades. Confounding by indication, unmeasured/residual confounding, outcome misclassification and immortal time bias are the most important biases. Although this should not be interpreted as an estimate of the risk of those biases, it may indicate which situations have raised greater interest among researchers so far and therefore should be especially considered in future studies using secondary databases to prevent their occurrence.

Appropriate methodological designs and application of statistical analysis techniques must be considered to control such situations. These strategies, summarized in Table 2, are also discussed in this review. In general, before initiating a research using secondary databases, researchers should assess in detail the sources of data available, focusing on the purpose for which they were created, and so become aware of their potential for bias. Medical records linkage with administrative databases can be useful to minimize the risk of bias, as well as the supplement or validation of secondary data with primary data (i.e. collected from ad hoc methods) when the completeness or quality of original data is questionable.

## Additional file

**Additional file 1:** Details of articles included in the review. (PDF 227 kb)

### Abbreviations

MeSH: Medical Subject Headings; PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses

### Acknowledgements

Not applicable.

### Funding

This work is funded by Grant ED431C 2018/20 from the Regional Ministry of Education, University and Vocational Training (*Consellería de Educación, Universidad y Formación Profesional, Xunta de Galicia*), Santiago de Compostela, Spain.

### Availability of data and materials

All data generated or analysed during this study are included in this published article.

### Authors' contributions

AF and GP-R contributed to study conception and design. All authors contributed to searching, screening, data collection and analyses. GP-R was responsible for drafting the manuscript. BT and AF provided comments and made several revisions of the manuscript. All authors read and approved the final version.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

<sup>1</sup>Department of Preventive Medicine and Public Health, University of Santiago de Compostela, c/ San Francisco s/n, 15786 Santiago de Compostela, A Coruña, Spain. <sup>2</sup>Health Research Institute of Santiago de Compostela (Instituto de Investigación Sanitaria de Santiago de Compostela - IDIS), Clinical University Hospital of Santiago de Compostela, 15706 Santiago de Compostela, Spain. <sup>3</sup>Consortium for Biomedical Research in Epidemiology & Public Health (CIBER en Epidemiología y Salud Pública - CIBERESP), Santiago de Compostela, Spain.

Received: 28 April 2018 Accepted: 26 February 2019

Published online: 11 March 2019

### References

- Hennessy S. Use of health care databases in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol.* 2006;98:311–3.
- Ray WA. Improving automated database studies. *Epidemiology.* 2011;22:302–4.
- 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.
- European Commission. Regulation (EU) 2016/679 of the European Parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation). 2016. <http://eur-lex.europa.eu/eli/reg/2016/679/oj>. Accessed 8 Oct 2018.
- U.S. Department of Health and Human Services. Code of Federal Regulations. Title 45 Public Welfare. Part 46 Protection of Human Subjects. 2016. <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html>. Accessed 8 Oct 2018.
- Moore TJ, Furberg CD. Electronic health data for postmarket surveillance: a vision not realized. *Drug Saf.* 2015;38:601–10.
- Gagne JJ. Restrictive reimbursement policies: bias implications for claims-based drug safety studies. *Drug Saf.* 2014;37:771–6.
- van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. *J Clin Epidemiol.* 2012;65:126–31.
- Weiss NS. The new world of data linkages in clinical epidemiology: are we being brave or foolhardy? *Epidemiology.* 2011;22:292–4.
- Colquhoun HL, Levac D, O'Brien KK, Straus S, Tricco AC, Perrier L, et al. Scoping reviews: time for clarity in definition, methods, and reporting. *J Clin Epidemiol.* 2014;67:1291–4.
- Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc.* 2015;13:141–6.
- Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med.* 2018;169:467–73.
- U.S. National Library of Medicine. Citations Added to MEDLINE® by Fiscal Year. 2016. [https://www.nlm.nih.gov/bsd/stats/cit\\_added.html](https://www.nlm.nih.gov/bsd/stats/cit_added.html). Accessed 8 Oct 2018.
- Kim DH, Schneeweiss S. Measuring frailty using claims data for pharmacoepidemiologic studies of mortality in older adults: evidence and recommendations. *Pharmacoepidemiol Drug Saf.* 2014;23:891–901.
- Schneeweiss S, Setoguchi S, Brookhart MA, Kaci L, Wang PS. Assessing residual confounding of the association between antipsychotic medications and risk of death using survey data. *CNS Drugs.* 2009;23:171–80.
- Strom BL. Methodologic challenges to studying patient safety and comparative effectiveness. *Med Care.* 2007;45(Suppl 2):S13–5.

17. Cohen JM, Wood ME, Hernandez-Diaz S, Nordeng H. Agreement between paternal self-reported medication use and records from a national prescription database. *Pharmacoepidemiol Drug Saf.* 2018;27:413–21.
18. Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med Care.* 2010;48(Suppl 1):S114–20.
19. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Stat Methods Med Res.* 2009;18:67–80.
20. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol.* 2003;158:280–7.
21. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol.* 2006;163:1149–56.
22. Smeeth L, Douglas I, Hubbard R. Commentary: we still need observational studies of drugs—they just need to be better. *Int J Epidemiol.* 2006;35:1310–1.
23. Joffe MM. Confounding by indication: the case of calcium channel blockers. *Pharmacoepidemiol Drug Saf.* 2000;9:37–41.
24. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf.* 2010;19:537–54.
25. Wittteman JC, D'Agostino RB, Stijnen T, Kannel WB, Cobb JC, de Ridder MA, et al. G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham heart study. *Am J Epidemiol.* 1998;148:390–401.
26. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11:550–60.
27. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf.* 2006;15:291–303.
28. Stürmer T, Glynn RJ, Rothman KJ, Avorn J, Schneeweiss S. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. *Med Care.* 2007;45:S158–65.
29. Lunt M, Glynn RJ, Rothman KJ, Avorn J, Stürmer T. Propensity score calibration in the absence of surrogacy. *Am J Epidemiol.* 2012;175:1294–302.
30. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol.* 2010;172:843–54.
31. Delaney JA, Suissa S. The case–crossover study design in pharmacoepidemiology. *Stat Methods Med Res.* 2009;18:53–65.
32. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nat Rev Rheumatol.* 2015;11:437–41.
33. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol.* 2016;183:758–64.
34. Zhang Y, Thamer M, Kaufman J, Cotter D, Hernán MA. Comparative effectiveness of two anemia management strategies for complex elderly dialysis patients. *Med Care.* 2014;52(Suppl 3):S132–9.
35. Glynn RJ, Knight EL, Levin R, Avorn J. Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology.* 2001;12:682–9.
36. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol.* 2007;166:646–55.
37. Shrank WH, Patrick AR, Brookhart MA. Healthy user and related biases in observational studies of preventive interventions: a primer for physicians. *J Gen Intern Med.* 2011;26:546–50.
38. Stürmer T, Jonsson Funk M, Poole C, Brookhart MA. Nonexperimental comparative effectiveness research using linked healthcare databases. *Epidemiology.* 2011;22:298–301.
39. Tamim H, Monfared AA, LeLorier J. Application of lag–time into exposure definitions to control for protopathic bias. *Pharmacoepidemiol Drug Saf.* 2007;16:250–8.
40. Patorno E, Garry EM, Patrick AR, Schneeweiss S, Gillet VG, Zorina O, et al. Addressing limitations in observational studies of the association between glucose–lowering medications and all–cause mortality: a review. *Drug Saf.* 2015;38:295–310.
41. Patorno E, Patrick AR, Garry EM, Schneeweiss S, Gillet VG, Bartels, et al. Observational studies of the association between glucose–lowering medications and cardiovascular outcomes: addressing methodological limitations. *Diabetologia.* 2014;57:2237–50.
42. Hernán MA, Hernández–Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology.* 2004;15:615–25.
43. Rebolj Kodre A, Pohar PM. Informative censoring in relative survival. *Stat Med.* 2013;32:4791–802.
44. Danaei G, Tavakkoli M, Hernán 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.
45. Ray WA. Evaluating medication effects outside of clinical trials: new–user designs. *Am J Epidemiol.* 2003;158:915–20.
46. Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Curr Epidemiol Rep.* 2015;2:221–8.
47. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. *Med Care.* 2013;51(Suppl 3):S80–6.
48. Rubin DB. Inference and missing data. *Biometrika.* 1976;63:581–92.
49. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. *Test (Madr).* 2009;18:1–43.
50. Moodie EE, Delaney JA, Lefebvre G, Platt RW. Missing confounding data in marginal structural models: a comparison of inverse probability weighting and multiple imputation. *Int J Biostat.* 2008;4:Article 13.
51. Siddique J, Harel O, Crespi CM. Addressing missing data mechanism uncertainty using multiple-model multiple imputation: application to a longitudinal clinical trial. *Ann Appl Stat.* 2012;6:1814–37.
52. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res.* 2013;22:278–9.
53. Resseguier N, Giorgi R, Paoletti X. Sensitivity analysis when data are missing not-at-random. *Epidemiology.* 2011;22:282.
54. de Groot MC, Klungel OH, Leufkens HG, van Dijk L, Grobbee DE, van de Garde EM. Sources of heterogeneity in case–control studies on associations between statins, ACE–inhibitors, and proton pump inhibitors and risk of pneumonia. *Eur J Epidemiol.* 2014;29:767–75.
55. Gamble JM, McAlister FA, Johnson JA, Eurich DT. Quantifying the impact of drug exposure misclassification due to restrictive drug coverage in administrative databases: a simulation cohort study. *Value Health.* 2012;15:191–7.
56. Kosinski AS, Flanders WD. Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: a regression approach. *Stat Med.* 1999;18:2795–808.
57. Ali A. Methodological challenges in observational research: a pharmacoepidemiological perspective. *Br J Pharm Res.* 2013;3:161–75.
58. Takahashi Y, Nishida Y, Asai S. Utilization of health care databases for pharmacoepidemiology. *Eur J Clin Pharmacol.* 2012;68:123–9.
59. Prada-Ramallal G, Takkouche B, Figueiras A. Summarising the evidence for drug safety: a methodological discussion of different meta-analysis approaches. *Drug Saf.* 2017;40:547–58.
60. Lin KJ, Schneeweiss S. Considerations for the analysis of longitudinal electronic health records linked to claims data to study the effectiveness and safety of drugs. *Clin Pharmacol Ther.* 2016;100:147–59.
61. Dokholyan RS, Muhlbaier LH, Falletta JM, Jacobs JP, Shahian D, Haan CK, et al. Regulatory and ethical considerations for linking clinical and administrative databases. *Am Heart J.* 2009;157:971–82.
62. Schneeweiss S, Wang PS. Claims data studies of sedative–hypnotics and hip fractures in older people: exploring residual confounding using survey information. *J Am Geriatr Soc.* 2005;53:948–54.
63. Haneuse S, Bogart A, Jazic I, Westbrook EO, Boudreau D, Theis MK, et al. Learning about missing data mechanisms in electronic health records–based research: a survey–based approach. *Epidemiology.* 2016;27:82–90.
64. Prada-Ramallal G, Takkouche B, Figueiras A. Diverging conclusions from the same meta-analysis in drug safety: source of data (primary versus secondary) takes a toll. *Drug Saf.* 2017;40:351–8.
65. Prada-Ramallal G, Roque F, Herdeiro MT, Takkouche B, Figueiras A. Primary versus secondary source of data in observational studies and heterogeneity in meta-analyses of drug effects: a survey of major medical journals. *BMC Med Res Methodol.* 2018;18:97.
66. Donahue JG, Weiss ST, Livingston JM, Goetsch MA, Greineder DK, Platt R. Inhaled steroids and the risk of hospitalization for asthma. *JAMA.* 1997;277:887–91.
67. Rochon PA, Tu JV, Anderson GM, Gurwitz JH, Clark JP, Lau P, et al. Rate of heart failure and 1-year survival for older people receiving low-dose beta-blocker therapy after myocardial infarction. *Lancet.* 2000;356:639–44.



68. Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ*. 2010;340:b5087.
69. Suissa S. Immortal time bias in pharmacoepidemiology. *Am J Epidemiol*. 2008;167:492–9.
70. Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol Drug Saf*. 2007;16:241–9.
71. Suissa S, Ernst P. Bias in observational study of the effectiveness of nasal corticosteroids in asthma. *J Allergy Clin Immunol*. 2005;115:714–9.
72. Suissa S. Inhaled steroids and mortality in COPD: bias from unaccounted immortal time. *Eur Respir J*. 2004;23:391–5.
73. Sin DD, Man SF, Tu JV. Inhaled glucocorticoids in COPD: immortal time bias. *Am J Respir Crit Care Med*. 2003;168:126–7.
74. Pride NB, Vestbo J, Soriano JB, Kiri VA. Inhaled glucocorticoids in COPD: immortal time bias. *Am J Respir Crit Care Med*. 2003;168:127.
75. Suissa S. Effectiveness of inhaled corticosteroids in chronic obstructive pulmonary disease: immortal time bias in observational studies. *Am J Respir Crit Care Med*. 2003;168:49–53.
76. Mantel N, Byar DP. Evaluation of response–time data involving transient states: an illustration using heart–transplant data. *J Am Stat Assoc*. 1974;69:81–6.
77. Kiri VA, Mackenzie G. Re: "immortal time bias in pharmacoepidemiology". *Am J Epidemiol*. 2009;170:667–8 author reply 668–9.
78. Karim ME, Gustafson P, Petkau J, Tremlett H. Long-term benefits and adverse effects of Beta-interferon for multiple sclerosis (BeAMS) study group. Comparison of statistical approaches for dealing with immortal time bias in drug effectiveness studies. *Am J Epidemiol*. 2016;184:325–35.
79. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. 2005;8:19–32.
80. Weinstein RB, Ryan P, Berlin JA, Matcho A, Schuemie M, Swerdel J, et al. Channeling in the use of nonprescription paracetamol and ibuprofen in an electronic medical records database: evidence and implications. *Drug Saf*. 2017;40:1279–92.
81. Pottegård A, Friis S, Stürmer T, Hallas J, Bahmanyar S. Considerations for pharmacoepidemiological studies of drug–cancer associations. *Basic Clin Pharmacol Toxicol*. 2018;122:451–9.
82. Melamed A, Rauh-Hain JA, Schorge JO. Clinical outcomes research in gynecologic oncology. *Gynecol Oncol*. 2017;146:653–60.
83. Dong YH, Alcusky M, Maio V, Liu J, Liu M, Wu LC, et al. Evidence of potential bias in a comparison of  $\beta$  blockers and calcium channel blockers in patients with chronic obstructive pulmonary disease and acute coronary syndrome: results of a multinational study. *BMJ Open*. 2017;7:e012997.
84. Bourbeau J, Aaron SD, Barnes NC, Davis KJ, Lacasse Y, Nadeau G. Evaluating the risk of pneumonia with inhaled corticosteroids in COPD: retrospective database studies have their limitations SA. *Respir Med*. 2017;123:94–7.
85. Macías Saint-Gerons D, de la Fuente HC, de Andrés TF, Catalá-López F. Future perspective of pharmacoepidemiology in the "big data era" and the growth of information sources. *Rev Esp Salud Publica*. 2016;90:e1–7.
86. Hudson M, Tascilar K, Suissa S. Comparative effectiveness research with administrative health data in rheumatoid arthritis. *Nat Rev Rheumatol*. 2016;12:358–66.
87. Bérard A, Wisner KL, Hultzsch S, Chambers C. Field studies versus database studies on the risks and benefits of medication use during pregnancy: distinct pieces of the same puzzle. *Reprod Toxicol*. 2016;60:123–8.
88. Haneuse S. Distinguishing selection bias and confounding bias in comparative effectiveness research. *Med Care*. 2016;54:e23–9.
89. Filion KB, Eberg M, Ernst P. Confounding by drug formulary restriction in pharmacoepidemiologic research. *Pharmacoepidemiol Drug Saf*. 2016;25:278–86.
90. Golozar A, Liu S, Lin JA, Peairs K, Yeh HC. Does metformin reduce cancer risks? Methodologic considerations. *Curr Diab Rep*. 2016;16:4.
91. Willis AW. Using administrative data to examine health disparities and outcomes in neurological diseases of the elderly. *Curr Neurol Neurosci Rep*. 2015;15:75.
92. Swanson SA, Hernandez-Diaz S, Palmsten K, Mogun H, Olfson M, Huybrechts KF. Methodological considerations in assessing the effectiveness of antidepressant medication continuation during pregnancy using administrative data. *Pharmacoepidemiol Drug Saf*. 2015;24:934–42.
93. Heinze G, Wallisch C, Kainz A, Hronsky M, Leffondré K, Oberbauer R, et al. Chances and challenges of using routine data collections for renal health care research. *Nephrol Dial Transplant*. 2015;30(Suppl 4):iv68–75.
94. Moulis G, Lapeyre-Mestre M, Palmaro A, Pugno G, Montastruc JL, Sailler L. French health insurance databases: what interest for medical research? *Rev Med Interne*. 2015;36:411–7.
95. Jensen ET, Cook SF, Allen JK, Logie J, Brookhart MA, Kappelman MD, et al. Enrollment factors and bias of disease prevalence estimates in administrative claims data. *Ann Epidemiol*. 2015;25:519–525.e2.
96. Funk MJ, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. *Curr Epidemiol Rep*. 2014;1:175–85.
97. Datta R, Kleinman K, Rifas-Shiman S, Placzek H, Lankiewicz J, Platt R, et al. Confounding by indication affects antimicrobial risk factors for methicillin-resistant *Staphylococcus aureus* but not vancomycin-resistant enterococci acquisition. *Antimicrob Resist Infect Control*. 2014;3:19.
98. Schneeweiss S. Learning from big health care data. *N Engl J Med*. 2014;370:2161–3.
99. Yang X, Chan JC. Metformin and the risk of cancer in type 2 diabetes: methodological challenges and perspectives. *Ann Transl Med*. 2014;2:52.
100. Zhang J, Curtis JR. Considerations in using registry and health plan data for studying pregnancy in rheumatic diseases. *Curr Opin Rheumatol*. 2014;26:315–20.
101. Gavriellov-Yusim N, Friger M. Use of administrative medical databases in population-based research. *J Epidemiol Community Health*. 2014;68:283–7.
102. Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. *AMIA Annu Symp Proc*. 2013;2013:1109–15.
103. Gallego B, Dunn AG, Coiera E. Role of electronic health records in comparative effectiveness research. *J Comp Eff Res*. 2013;2:529–32.
104. Ryan PB, Madigan D, Stang PE, Schuemie MJ, Hripscak G. Medication-wide association studies. *CPT Pharmacometrics Syst Pharmacol*. 2013;2:e76.
105. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51(8 Suppl 3):S30–7.
106. Suissa S, Azoulay L. Metformin and the risk of cancer: time-related biases in observational studies. *Diabetes Care*. 2012;35:2665–73.
107. Yang X, Weng J. Increased cancer risk with drug use among patients with diabetes: are the biased methods the culprit? *J Diabetes Investig*. 2012;3:479–80.
108. Hershman DL, Wright JD. Comparative effectiveness research in oncology methodology: observational data. *J Clin Oncol*. 2012;30:4215–22.
109. Suling M, Pigeot I. Signal detection and monitoring based on longitudinal healthcare data. *Pharmaceutics*. 2012;4:607–40.
110. Kiri VA. A pathway to improved prospective observational post-authorization safety studies. *Drug Saf*. 2012;35:711–24.
111. Yang XL, Ma RC, So WY, Kong AP, Xu G, Chan JC. Addressing different biases in analysing drug use on cancer risk in diabetes in non-clinical trial settings—what, why and how? *Diabetes Obes Metab*. 2012;14:579–85.
112. Nelson JC, Cook AJ, Yu O, Dominguez C, Zhao S, Greene SK, et al. Challenges in the design and analysis of sequentially monitored postmarket safety surveillance evaluations using electronic observational health care data. *Pharmacoepidemiol Drug Saf*. 2012;21(Suppl 1):62–71.
113. Fung V, Brand RJ, Newhouse JP, Hsu J. Using medicare data for comparative effectiveness research: opportunities and challenges. *Am J Manag Care*. 2011;17:488–96.
114. Zhang J, Yun H, Wright NC, Kilgore M, Saag KG, Delzell E. Potential and pitfalls of using large administrative claims data to study the safety of osteoporosis therapies. *Curr Rheumatol Rep*. 2011;13:273–82.
115. Hernan MA. With great data comes great responsibility: publishing comparative effectiveness research in epidemiology. *Epidemiology*. 2011;22:290–1.
116. McNeil JJ, Piccenna L, Ronaldson K, Ioannides-Demos LL. The value of patient-centred registries in phase IV drug surveillance. *Pharmaceut Med*. 2010;24:281–8.
117. Ehrenstein V, Sørensen HT, Bakketeig LS, Pedersen L. Medical databases in studies of drug teratogenicity: methodological issues. *Clin Epidemiol*. 2010;2:37–43.
118. Hudson M, Suissa S. Avoiding common pitfalls in the analysis of observational studies of new treatments for rheumatoid arthritis. *Arthritis Care Res (Hoboken)*. 2010;62:805–10.
119. Martin-Latry K, Bégaud B. Pharmacoepidemiological research using French reimbursement databases: yes we can! *Pharmacoepidemiol Drug Saf*. 2010;19:256–65.

120. Giezen TJ, Mantel-Teeuwisse AK, Leufkens HG. Pharmacovigilance of biopharmaceuticals: challenges remain. *Drug Saf.* 2009;32:811–7.
121. Harpe SE. Using secondary data sources for pharmacoepidemiology and outcomes research. *Pharmacotherapy.* 2009;29:138–53.
122. Giordano SH, Kuo YF, Duan Z, Hortobagyi GN, Freeman J, Goodwin JS. Limits of observational data in determining outcomes from cancer therapy. *Cancer.* 2008;112:2456–66.
123. Pigeot I, Ahrens W. Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations. *Pharmacoepidemiol Drug Saf.* 2008;17:215–23.
124. Cramer JA, Silverman SL, Gold DT. Methodological considerations in using claims databases to evaluate persistence with bisphosphonates for osteoporosis. *Curr Med Res Opin.* 2007;23:2369–77.
125. Terris DD, Litaker DG, Koroukian SM. Health state information derived from secondary databases is affected by multiple sources of bias. *J Clin Epidemiol.* 2007;60:734–41.
126. Hughes MD, Williams PL. Challenges in using observational studies to evaluate adverse effects of treatment. *N Engl J Med.* 2007;356:1705–7.
127. de Vries F, de Vries C, Cooper C, Leufkens B, van Staa TP. Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: the general practice research database. *Int J Epidemiol.* 2006;35:1301–8.
128. Suissa S. Observational studies of inhaled corticosteroids in chronic obstructive pulmonary disease: misconstrued immortal time bias. *Am J Respir Crit Care Med.* 2006;173:464 author reply 464–5.
129. Etminan M, Gill S, Fitzgerald M, Samii A. Challenges and opportunities for pharmacoepidemiology in drug-therapy decision making. *J Clin Pharmacol.* 2006;46:6–9.
130. Ray WA. Observational studies of drugs and mortality. *N Engl J Med.* 2005; 353:2319–21.
131. Holbrook A, Grootendorst P, Willison D, Goldsmith C, Sebaldt R, Keshavjee K. Can current electronic systems meet drug safety and effectiveness requirements? *AMIA Annu Symp Proc.* 2005;335–9.
132. Schneeweiss S, Wang PS. Association between SSRI use and hip fractures and the effect of residual confounding bias in claims database studies. *J Clin Psychopharmacol.* 2004;24:632–8.
133. Kiri VA, Vestbo J, Pride NB, Soriano JB. Inhaled steroids and mortality in COPD: bias from unaccounted immortal time. *Eur Respir J.* 2004;24:190–1 author reply 191–2.
134. Bannwarth B. Gastrointestinal safety of paracetamol: is there any cause for concern? *Expert Opin Drug Saf.* 2004;3:269–72.
135. Zhan C, Miller MR. Administrative data based patient safety research: a critical review. *Qual Saf Health Care.* 2003;12(Suppl 2):ii58–63.
136. Strom BL. Data validity issues in using claims data. *Pharmacoepidemiol Drug Saf.* 2001;10:389–92.
137. Sørensen HT, Johnsen SP, Nørgård B. Methodological issues in using prescription and other databases in pharmacoepidemiology. *Nor Epidemiol.* 2001;11:13–8.
138. Hallas J. Pharmacoepidemiology – current opportunities and challenges. *Nor Epidemiol.* 2001;11:7–12.
139. Skegg DC. Pitfalls of pharmacoepidemiology. *BMJ.* 2000;321:1171–2.
140. Cepeda MS, Fife D, Denarié M, Bradford D, Roy S, Yuan Y. Quantification of missing prescriptions in commercial claims databases: results of a cohort study. *Pharmacoepidemiol Drug Saf.* 2017;26:386–92.
141. Camplain R, Kucharska-Newton A, Cuthbertson CC, Wright JD, Alonso A, Heiss G. Misclassification of incident hospitalized and outpatient heart failure in administrative claims data: the atherosclerosis risk in communities (ARIC) study. *Pharmacoepidemiol Drug Saf.* 2017;26:421–8.
142. Palmaro A, Moulis G, Despas F, Dupouy J, Lapeyre-Mestre M. Overview of drug data within French health insurance databases and implications for pharmacoepidemiological studies. *Fundam Clin Pharmacol.* 2016;30:616–24.
143. Lanes S, Brown JS, Haynes K, Pollack MF, Walker AM. Identifying health outcomes in healthcare databases. *Pharmacoepidemiol Drug Saf.* 2015;24: 1009–16.
144. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data resource profile: clinical practice research datalink (CPRD). *Int J Epidemiol.* 2015;44:827–36.
145. Weil G, Motamed C, Eghiaian A, Guye ML, Bourgain JL. The use of a clinical database in an anesthesia unit: focus on its limits. *J Clin Monit Comput.* 2015;29:163–7.
146. Li X, Stürmer T, Brookhart MA. Evidence of sample use among new users of statins: implications for pharmacoepidemiology. *Med Care.* 2014;52:773–80.
147. Lauffenburger JC, Balasubramanian A, Farley JF, Critchlow CW, O'Malley CD, Roth MT, et al. Completeness of prescription information in US commercial claims databases. *Pharmacoepidemiol Drug Saf.* 2013;22:899–906.
148. Maciejewski ML. Potential bias in medication adherence studies of prevalent users. *Health Serv Res.* 2013;48:1468–86.
149. Grimes DA. Epidemiologic research using administrative databases: garbage in, garbage out. *Obstet Gynecol.* 2010;116:1018–9.
150. Velthove KJ, Leufkens HG, Souverein PC, Schweizer RC, van Solinge WW. Testing bias in clinical databases: methodological considerations. *Emerg Themes Epidemiol.* 2010;7:2.
151. Oostenbrink R, Moons KG, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. *J Clin Epidemiol.* 2003;56:501–6.
152. Wade RL, Patel JG, Hill JW, De AP, Harrison DJ. Estimation of missed statin prescription use in an administrative claims dataset. *J Manag Care Spec Pharm.* 2017;23:936–42.
153. Czwikla J, Jobski K, Schink T. The impact of the lookback period and definition of confirmatory events on the identification of incident cancer cases in administrative data. *BMC Med Res Methodol.* 2017;17:122.
154. Pauly NJ, Talbert JC, Brown J. Low-cost generic program use by Medicare beneficiaries: implications for medication exposure misclassification in administrative claims data. *J Manag Care Spec Pharm.* 2016;22:741–51.
155. Mazzali C, Paganoni AM, Ieva F, Masella C, Maistrillo M, Agostoni O, et al. Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy region, 2000 to 2012. *BMC Health Serv Res.* 2016;16:234.
156. Hampf C, Greene P, Pinheiro SP. Use of prescription drug samples in the USA: a descriptive study with considerations for pharmacoepidemiology. *Drug Saf.* 2016;39:261–70.
157. Winterstein AG, Kubilis P, Bird S, Cooper-DeHoff RM, Nichols GA, Delaney JA. Misclassification in assessment of diabetogenic risk using electronic health records. *Pharmacoepidemiol Drug Saf.* 2014;23:875–81.
158. Skurtveit S, Selmer R, Tverdal A, Furu K, Nystad W, Handal M. Drug exposure: inclusion of dispensed drugs before pregnancy may lead to underestimation of risk associations. *J Clin Epidemiol.* 2013;66:964–72.
159. Gamble JM, McAlister FA, Johnson JA, Eurich DT. Restrictive drug coverage policies can induce substantial drug exposure misclassification in pharmacoepidemiologic studies. *Clin Ther.* 2012;34:1379–86.
160. van Walraven C, Bennett C, Forster AJ. Administrative database research infrequently used validated diagnostic or procedural codes. *J Clin Epidemiol.* 2011;64:1054–9.
161. Hoover KW, Tao G, Kent CK, Aral SO. Epidemiologic research using administrative databases: garbage in, garbage out. *Obstet Gynecol.* 2011; 117:729 author reply 729–30.
162. Dore DD, Chaudhry S, Hoffman C, Seeger JD. Stratum-specific positive predictive values of claims for acute pancreatitis among commercial health insurance plan enrollees with diabetes mellitus. *Pharmacoepidemiol Drug Saf.* 2011;20:209–13.
163. Lanes SF, de Luise C. Bias due to false-positive diagnoses in an automated health insurance claims database. *Drug Saf.* 2006;29:1069–75.
164. Ray WA, Thapa PB, Gideon P. Misclassification of current benzodiazepine exposure by use of a single baseline measurement and its effects upon studies of injuries. *Pharmacoepidemiol Drug Saf.* 2002;11:663–9.
165. Palmaro A, Boucherie Q, Dupouy J, Micallef J, Lapeyre-Mestre M. Immeasurable time bias due to hospitalization in medico-administrative databases: which impact for pharmacoepidemiological studies? *Pharmacoepidemiol Drug Saf.* 2017;26:544–53.
166. Targownik LE, Suissa S. Understanding and avoiding immortal-time bias in gastrointestinal observational research. *Am J Gastroenterol.* 2015;110:1647–50.
167. Matok I, Azoulay L, Yin H, Suissa S. Immortal time bias in observational studies of drug effects in pregnancy. *Birth Defects Res A Clin Mol Teratol.* 2014;100:658–62.
168. Cook EA, Schneider KM, Chrischilles E, Brooks JM. Accounting for unobservable exposure time bias when using medicare prescription drug data. *Medicare Medicaid Res Rev.* 2013;3.
169. Suissa S, Dell'aniello S, Vahey S, Renoux C. Time-window bias in case-control studies: statins and lung cancer. *Epidemiology.* 2011;22:228–31.
170. Suissa S. Immeasurable time bias in observational studies of drug effects on mortality. *Am J Epidemiol.* 2008;168:329–35.