


REVIEW

A systematic review of how missing data are handled and reported in multi-database pharmacoepidemiologic studies

Nicholas B. Hunt¹  | Helga Gardarsdottir^{1,2,3}  | Marloes T. Bazelier¹ |
Olaf H. Klungel¹ | Romin Pajouheshnia¹ 

¹Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, The Netherlands

²Department of Clinical Pharmacy, University Medical Centre Utrecht, Utrecht, The Netherlands

³Department of Pharmaceutical Sciences, School of Health Sciences, University of Iceland, Reykjavik, Iceland

Correspondence

Romin Pajouheshnia, Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, David de Wiedgebouw, Universiteitsweg 99, 3584 CG Utrecht, The Netherlands.
Email: r.pajouheshnia@uu.nl

Abstract

Purpose: Pharmacoepidemiologic multi-database studies (MDBS) provide opportunities to better evaluate the safety and effectiveness of medicines. However, the issue of missing data is often exacerbated in MDBS, potentially resulting in bias and precision loss. We sought to measure how missing data are being recorded and addressed in pharmacoepidemiologic MDBS.

Methods: We conducted a systematic literature search in PubMed for pharmacoepidemiologic MDBS published between 1st January 2018 and 31st December 2019. Included studies were those that used ≥ 2 distinct databases to assess the same safety/effectiveness outcome associated with a drug exposure. Outcome variables extracted from the studies included strategies to execute a MDBS, reporting of missing data (type, bias evaluation) and the methods used to account for missing data.

Results: Two thousand seven hundred and twenty-six articles were identified, and 62 studies were included: using data from either North America (56%), Europe (31%), multiple regions (11%) or East-Asia (2%). Thirty-five (56%) articles reported missing data: 11 of these studies reported that this could have introduced bias and 19 studies reported a method to address missing data. Thirteen (68%) carried out a complete case analysis, 2 (11%) applied multiple imputation, 2 (11%) used both methods, 1 (5%) used mean imputation and 1 (5%) substituted information from a similar variable.

Conclusions: Just over half of the recent pharmacoepidemiologic MDBS reported missing data and two-thirds of these studies reported how they accounted for it. We should increase our vigilance for database completeness in MDBS by reporting and addressing the missing data that could introduce bias.

KEYWORDS

missing data, multi-database, pharmacoepidemiology, review

1 | INTRODUCTION

The use of multiple health databases in pharmacoepidemiologic studies can facilitate more robust assessments of drug safety and effectiveness.^{1,2} Multi-database studies (MDBS) involve the analysis of routinely collected data from two or more data sources, which may take the form of health insurance claims databases, electronic healthcare records (EHR) or healthcare record linkage systems.³ MDBS can allow us to investigate specific subgroups of patients, rare outcomes, or conduct an early post-approval assessment of safety and effectiveness.^{4,5} Multi-national MDBS can lead to more generalisable results, due to the inclusion of heterogeneous patient populations⁴ and allow us to compare the safety and effectiveness of compounds between countries and regions, taking differences in ethnicity and health care systems into account.⁵

MDBS are methodologically more complex than single database studies as a result of inter-database heterogeneity, caused by differences in what information is recorded and how. This brings practical challenges such as how to standardise analyses across a distributed network and how to combine data or results.^{2,6,7} In addition, there may also be differences in the completeness of information across databases, potentially resulting in missing data, herein defined as any data that are relevant to the analysis or interpretation of a study that are not available to the study investigators at the time of analysis or reporting. Common data models (CDMs) and common protocols (CPs) have been used across database networks to mitigate bias due to heterogeneity in MDBS analyses,² but cannot solve differences in database completeness. Failure to account for missing data in epidemiologic studies can introduce bias, even altering the direction of treatment effect estimates, and reduce the precision of effect estimates.^{8,9} For example, one study showed that risk estimates of venous thromboembolism associated with anti-osteoporotic medications were substantially affected by the use of different strategies for the handling of missing data, leading to differences in the direction of treatment effect estimates.⁸

Missing data can arise at several stages within a multi-database pharmacoepidemiologic study. Like in a single database study, data may not be recorded at the stage of data entry into the database. For example, data may be partially (or sporadically) missing because a health professional recorded information in an unstructured manner (e.g., using free text) or did not collect information for certain patients, such as body mass index (BMI) and smoking status, because they were considered healthy.¹⁰⁻¹² MDBS can have the additional complexity of completely (or systematically) missing variables, where data on a certain variable are missing for all individuals in a database.^{12,13} This may occur when certain variables are not recorded in a database because they are not required by the health authorities for reimbursement (in administrative/claims databases) or because the recording is not part of routine clinical practice (in electronic health record (EHR) databases). In the case of a systematically missing confounder, this can lead to residual confounding in a study; if a variable used to define exposure is systematically missing, this can lead to exposure misclassification in one or more of the databases. There may also be some information loss during the extract, transform and load process

Key Points

- Missing data can lead to biased estimates of drug safety and effectiveness, a problem that can be exacerbated in multi-database studies.
- Forty-four percent of recent multi-database pharmacoepidemiologic studies did not report missing data and 69% did not report accounting for missing data.
- In studies which report missing data, lifestyle variables were most frequently reported missing (14%–29%).
- Most studies which reported a method to address missing data performed a complete case analysis (68%). Multiple imputation was the predominant statistical method used to handle missing data (22%).
- Variables with missing data, the potential bias and accounting for missing data should be thoroughly reported.

(e.g., data which did not meet the quality criteria are omitted) or creation of final analytical variables. This can happen, for example, when components of a composite variable are missing or if time restrictions are applied to a variable, such as a measurement being available within 12 months of the index date.¹⁴

Methods to account for sporadically missing data, such as multiple imputation (MI) and inverse probability weighting, are widely known.^{8,15} To handle systematically missing data, a practical approach is to exclude the missing variable from the analyses or exclude an entire database.⁸ A recently proposed alternative is multi-level MI (MLMI), which can account for both sporadically and systematically missing data. This approach utilises information on the covariance of variables in one dataset or database to impute missing information in another.¹⁶⁻¹⁹

The reporting of missing data and the methods used to address it are specified in the RECORD-PE and STROBE statements.^{20,21} Without thorough reporting, we cannot have full confidence in the validity of the study estimates. Our aim for this review was to first measure the extent to which recent MDBS reported missing data and considered it as a potential source of bias. Second, we sought to identify which strategies are being reported in MDBS to deal with missing data.

2 | METHODS

2.1 | Study design and search criteria

We conducted a systematic literature search in PubMed to identify and report the methods used in recent peer-reviewed, multi-database pharmacoepidemiologic studies. The full list of inclusion and exclusion criteria can be seen in Table 1. The search strategy was adapted from a previous review of MDBS by Bazelier et al.¹ (see supplementary table 1 for the PubMed search terms). We restricted

TABLE 1 The selection criteria for published pharmacoepidemiologic multi-database studies to be included in the systematic review

Inclusion criteria	Exclusion criteria
Multi-database (of independent patient populations)	Spontaneous reporting databases
Observational (non-randomised) study	Methodological studies
Safety or effectiveness of a pharmaceutical compound	Drug utilisation studies
Publish date 01-01-2018 to 31-12-2019	Reviews
Peer-reviewed literature	Clinical trials
	Literature which is not published in English
	Non-routine data collection where data were collected for research purposes

the search to studies published between 01-01-2018 and 31-12-2019. This search was performed on 07-11-2019 for the studies up until 31-10-2019. The search was updated on 08-01-2020 to include the studies between 01-11-2019 and 31-12-2019.

We additionally performed a search of established database networks. These networks were identified by the authors, with the assistance of an external expert. The networks included those published by the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP), the US Food and Drug Administration's (FDA) Sentinel Initiative, the Canadian Network for Observational Drug Effect Studies (CNODES), the Asian Pharmacoepidemiology Network (AsPEN), the National Patient-Centered Clinical Research Network (PCORnet) and the Vaccine Safety Datalink (VSD).²²⁻²⁷ Where the network websites were not up-to-date with a full list of associated publications, the names of the networks were searched in PubMed (23-01-2020).

2.2 | Screening and selection

Title and abstract screening for eligibility were performed by one author (NBH). Another author (RP) independently screened the title and abstract of 100 articles. Any differences in this pilot screening were discussed and resolved. One author (NBH) then screened the full-text of the remaining articles for eligibility. Two authors (NBH and RP) independently reviewed the full list of included studies to confirm eligibility and disagreements were discussed with the other authors.

2.3 | Extraction of the general study characteristics

The general characteristics of each study were recorded: The study design (e.g., cohort or case-control), the journal, the exposure and

outcome, the size (number of databases, countries, subjects), the database type (administrative/claims, EHR, other) and whether the study provided a pooled estimate. To categorise the strategies used to execute the MDDBS, we used criteria described by ENCePP (a full description can be found in Gini et al.²⁸). We categorised each study as carrying out either a local analysis, where the data extraction and analysis are conducted by individual centres (according to a CP); sharing of raw data, where the local site extracts the raw data and transfers it to a central partner for the analysis; the use of a study-specific CDM; or use of a general CDM.²⁸

2.4 | Extraction of the outcomes

For the primary outcome, we measured how many of the included studies reported the existence of missing data; in what context this was reported (methods, results or as a limitation); which variables were missing data; type of missingness (sporadic or systematic, as determined by the authors); the type of variable with missing data (exposure, outcomes or confounders); and the amount of % missing. We recorded whether the authors of each study discussed the extent to which the missing data had contributed to bias and whether missing data were missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR).²⁹

For studies that reported missing data, we recorded which method was used to account for missing data. For studies which used MI, we additionally extracted information on the number of imputations per variable, the number of imputed variables and the statistical software used for the imputation.

2.5 | Data analysis

Data extraction was piloted in five articles by two authors (NBH and RP). The remaining data extraction was conducted by a single author (NBH) and any uncertainties were discussed with the co-authors. Extracted data were recorded and where applicable, data were transformed into a pre-specified answer list. Means, medians and ranges were calculated.

3 | RESULTS

The search identified 2726 publications for title/abstract review (Figure 1). Sixty-two articles from forty-four scientific journals were eligible for inclusion (see supplementary table 2 for a list of included articles). An overview of the general study characteristics can be seen in Table 2. Thirty-five (56%) of the included studies used exclusively North American data, 19 (31%) used exclusively European data, 7 (11%) used data from a combination of regions and 1 (2%) used exclusively East-Asian data (Table 2 and supplementary figure 1).

Fifty-seven (92%) of the included studies provided a pooled estimate of multiple databases in their analysis. Twenty-two (35%) of the

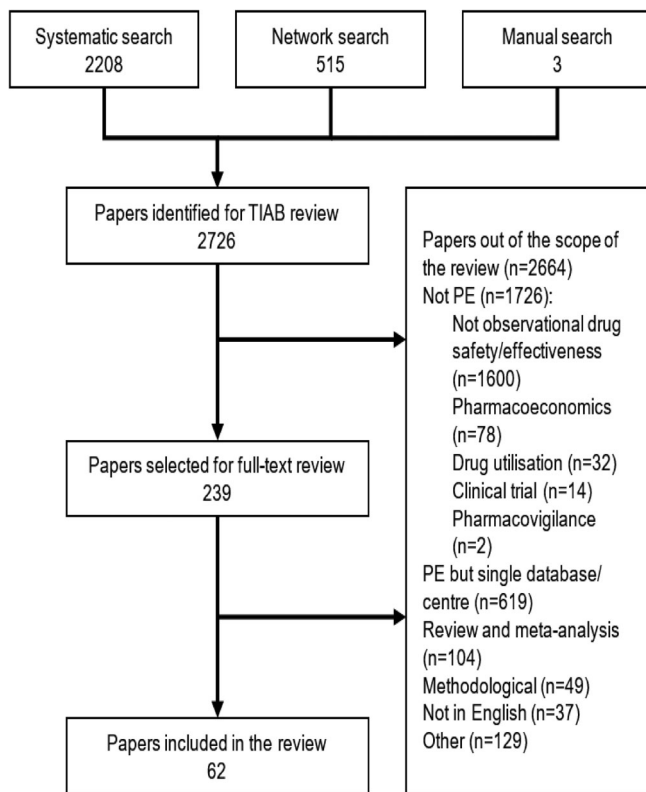


FIGURE 1 A flow diagram showing the selection process for the papers included and excluded from this systematic review. Other ($n = 129$): Presentations (62), case reports (29), not in human studies (8), letter/ comments (8), full-text unavailable (7), guidelines (4), duplicate (4), protocols (2), conference abstracts (1), book chapters (1), not in timeframe (1), and strategic plans (1). TAIB, title or abstract; PE, pharmacoepidemiology

included studies carried out their analysis locally per study site, of which five did not provide a pooled estimate of multiple databases. Ten (16%) reported sharing raw data for a central analysis and 24 (38%) used a CDM. Twenty-three (96%) of these used a general CDM with 11 using the VSD and 7 using Sentinel's CDM; one study used a study-specific CDM.

Missing data were reported in 35 (56%) articles, with potential confounders such as lifestyle factors reported missing most often: BMI ($n = 10$, 29%), smoking status ($n = 9$, 26%), age or date of birth ($n = 7$, 20%), ethnicity/race ($n = 6$, 17%) and education level ($n = 5$, 14%). The extent of the reporting varied greatly and in six cases missing data were inferred from close inspection of the text or tables. Five articles reported the percentage of missing data, either by reporting the percentage per variable or per individuals in a database with data missing on one or more variable. The amount of missing data (per variable or database) reported in studies ranged from <1% to 56%. For the other 27 (44%) studies, no evidence of missing data was reported. In those that used a CDM ($n = 24$), 58% reported missing data compared to 55% of studies which did not use a CDM ($n = 38$). Missing data were less often reported in studies which used North-American (54%) and European data (53%) compared to those which used data from a combination of regions (86%). The reporting of missing data

TABLE 2 General characteristics of the included multi-database pharmacoepidemiologic studies

Study characteristics		
<i>Safety or effectiveness</i>	<i>N (62)</i>	<i>%</i>
Safety	50	80
Effectiveness	6	10
Both	6	10
<i>Study design</i>	<i>N (62)</i>	<i>%</i>
Cohort	53	85
Case-control	9	15
<i>Exposure ATC anatomical group</i>	<i>N (62)</i>	<i>%</i>
Anti-infectives for systemic use	17	27
Alimentary tract and metabolism	10	16
Antineoplastic and immunomodulating agents	8	13
Cardiovascular system	6	10
Blood and blood-forming organs	6	10
Nervous system	5	8
Genitourinary system and sex hormones	2	3
Multiple groups	5	8
Other	3	5
<i>Database type</i>	<i>N (62)</i>	<i>%</i>
Administrative/claims	28	45
Electronic health records	12	19
Other	4	6
Combination	18	29
<i>Study size (individuals)</i>	<i>Mean</i>	<i>Range</i>
Cohort	324 190	568–6 177 795
Case-cohort	363 567	193–1 706 113
No. of databases	5	2–17
No. of countries	2	1–9
<i>MDBS strategies</i>		
<i>Pooled estimate</i>	<i>N (62)</i>	<i>%</i>
Yes	57	92
No	5	8
<i>MDBS strategy (ENCePP criteria)</i>	<i>N (62)</i>	<i>%</i>
Sharing of raw data	10	16
General CDM	23	37
Study-specific CDM	1	2
Local analysis	22	35
Not reported	6	10

Abbreviations: ATC, anatomical therapeutic chemical; MDBS, multi-database study; CDM, common data model; ENCePP, European Network of Centres for Pharmacoepidemiology and Pharmacovigilance.

differed between studies which used EHR data (67%), admin/claims databases (54%), a combination of EHR and admin/claims data (50%) and other types of databases such as registries (75%).

In the 35 studies which reported missing data, 13 (37%) reported the presence of sporadically missing data only, 4 (11%) reported the presence of systematically missing data only, 11 (31%) reported the presence of both and 7 (20%) were uncertain. Nine (26%) reported missing data in both the results and limitations sections, the rest reported in either section or not at all (Table 3). Fifteen articles (43%) reported that missing data existed within the constituent data sets without stating the quantity, while 14 (40%) reported the quantity of missing data as a percentage or number. Eleven (31%) studies reported that this missing data could

have introduced bias to their findings, but no studies reported the possible missingness mechanism in terms of MCAR, MAR or MNAR.

In the 19 studies that reported a method to address missing, 13 (68%) carried out a complete case analysis, 2 (11%) used MI separately per database, 2 (11%) applied and compared a complete case analysis and MI, 1 (5%) used mean imputation and 1 (5%) substituted the missing information with information from a similar variable, where missing data points are replaced with the 'best alternative'. For example, gestational age was determined by the first day of the last menstrual period if an ultrasound examination was missing. The studies which used both CCA and MI did so to test whether the separate strategies altered the overall outcome. For those that carried out MI, two studies reported the use of the Markov chain Monte Carlo method and two used MI by chained equations (MICE). Within one study which used MI, single imputation of the mode of the variable was used to address variables with <2% missing data. Studies imputed six ($n = 2$), four ($n = 1$), two ($n = 1$) or an unreported ($n = 1$) number of variables. The number of imputations per variable in the included studies which used MI were 20 ($n = 2$), 10 ($n = 2$) or unreported ($n = 1$). These imputations were performed in statistical software Stata ($n = 2$), R ($n = 1$), SAS ($n = 1$) or unreported ($n = 1$).

TABLE 3 An overview the reporting of missing data and missing data methods in the 62 recently published multi-database pharmacoepidemiology studies included in the systematic review

Reported missing data	N (62)	%
Yes	35	56
<i>Location:</i>		
In analysis	22	
In limitations	4	
In both	9	
<i>Variable type:</i>		
Exposure	5	
Outcome	8	
Confounder	31	
No	27	44
Evaluated missing data for bias		
Yes	11	18
No	51	82
Type of missingness		
Sporadic only	13	21
Systematic only	4	6
Both reported	11	18
Both uncertain	34	55
Types of missing variables reported		
Categorical only	11	18
Continuous only	6	10
Both	13	21
N/A or uncertain	32	52
Missing data method reported		
Yes	19	31
No	43	69
Missing data method used		
Complete case analysis (CCA)	13	21
Multiple imputation (MI)	2	3
Mean imputation	1	2
Substitution	1	2
Both MI and CCA	2	3
None or uncertain	43	69

4 | DISCUSSION

In this systematic review of recently published multi-database pharmacoepidemiologic studies, we found that out of 62 included articles, only 56% reported missing data, 18% reported whether missing data could have biased the study and 31% reported how they dealt with missing data. The reporting of missing data was slightly higher in studies which used a CDM and those which used European and Asian data compared to North-American data.

In contrast to Rioux et al.,³⁰ which focused primarily on survey-based observational research, instead of MDDBS, we found a substantially lower proportion of studies reported missing data (56% compared to 87%), reported a method to address missing data (31% compared to 77%) and the possible mechanisms behind the missingness pattern (0% compared to 11%). However, the comparison with survey-based observational research is not straightforward as we focus exclusively on database studies.

The inconsistent reporting of information about missing data observed in this review may also reflect that the completeness of information in healthcare databases varies according to the type of database and the type of information captured. EHR and health administrative databases are, by design, not created for the post-marketing assessment of medicines, thus desired information for research, particularly on potential confounders, may not be recorded.³ This may be in contrast to registries where information that is not routinely collected in clinical practice is gathered through forms or surveys. In databases where there are certain elements which could be considered complete, such as information on pharmacy dispensing in

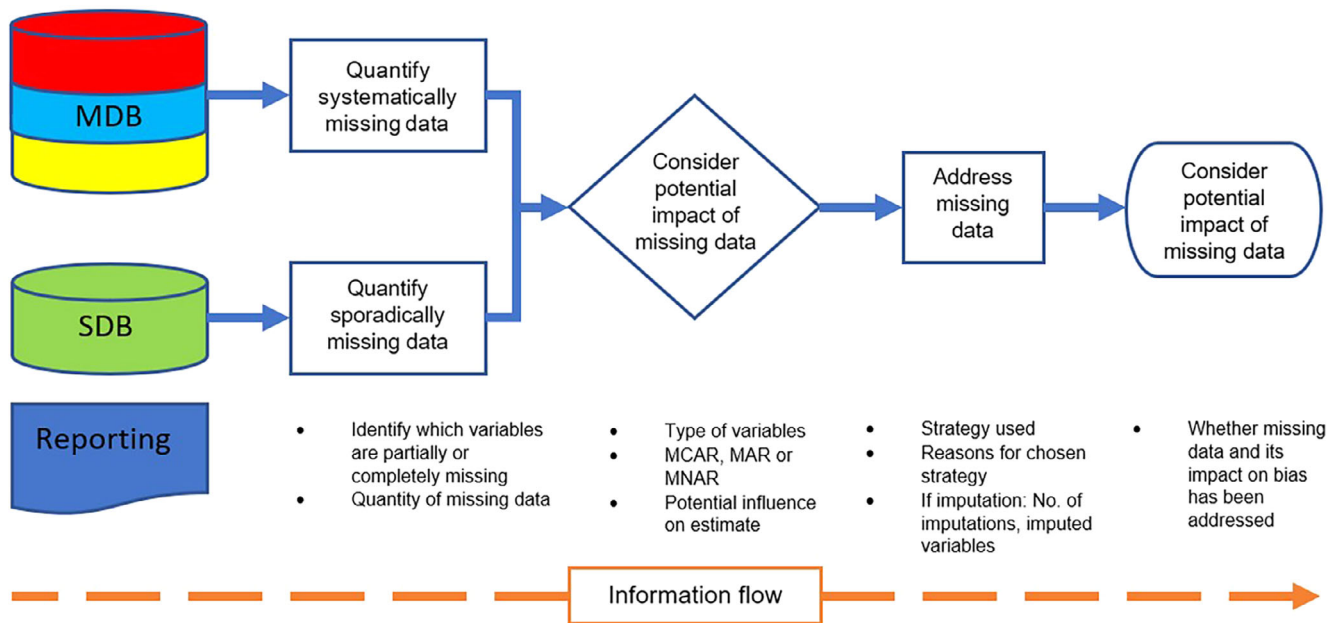


FIGURE 2 A flow diagram of the suggested process for reporting missing data and the use of missing data methods when publishing a multi-database pharmacoepidemiologic study. Basic details of missing data and missing data method which should be recorded and at which stage in a multi-database pharmacoepidemiologic study. MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random; MDB, multiple databases; SDB, single database [Colour figure can be viewed at wileyonlinelibrary.com]

databases used for reimbursement purposes; there still could be completely missing variables, such as lifestyle factors.

In MDBS which use claims databases, the issue of systematically missing data may be larger than that of sporadically missing data. In a pharmacoepidemiologic MDBS, there is a high likelihood of missing data because the multiple databases involved may record different variables. Furthermore, a recent study of established European health databases showed that vaccinations are captured in 38% of databases, while inpatient administered (5.8%) and over-the-counter drugs are rarely captured.³ It is therefore possible that the studies in this review have underreported missing data in their analyses, although this depends on the relevance of these kinds of exposures to the included studies.

The reporting of whether missing data could have contributed to potential bias is important, however, we find that only 21 (34%) studies did this. None of the included studies reports whether the data were missing at random (MAR, MCAR) or not (MNAR). This is a valuable factor when determining whether missing data have contributed to bias in the study and when considering what method to apply to correct for missing data. MNAR assumes that the missingness pattern is dependent on unobserved variables, so it is a potentially greater source of bias.⁹ Only 31% of the studies reported a missing data method, the other studies (69%) may have applied a missing data method without reporting it in the publication. In the absence of any additional missing data methods, a CCA is likely to have been performed. In 2012, it was reported that 81% of epidemiologic studies carry out a CCA.³¹ More recently, it was reported that 70% carry out CCA and 18%, MI.³⁰ It has already been recommended that missing data assumptions and the rationale for using a CCA should be reported, since CCA may bias the study estimates as data is often not

MCAR.^{8,9} Three of the included studies reported a head-to-head comparison: two found that the use of MI compared to CCA did not change the conclusions of the study and one reported difficulty in comparing the methods due to a large quantity of missing data. We recommend that future studies conduct and report sensitivity analyses to clarify the potential impact of methodological choices when addressing missing data in their analyses.⁸

One of the other possible solutions to deal with missing data is MI, which allows the inclusion of the patients who have missing data. It is a method that assumes data are MAR but it can handle data which is MCAR or (under stronger assumptions) MNAR.¹⁵ MI can improve precision in the study estimate compared to CCA, although its use did not make a substantial impact on the point estimate in the studies which reported using both methods separately.³²⁻³⁴ Since there was heterogeneity in the databases used, different MI models were applied within the studies to account for the available variables. None of the studies captured in our review applied a statistical method to address systematically missing information. Theoretically, imputation techniques can be applied to the pooled raw data of multiple databases, however, practically this is often not possible due to restrictions on the sharing of data. The use of MI also incorrectly assumes homogeneity of associations amongst (possibly) heterogeneous populations and data sources.¹² MI can potentially be applied to multi-level data structures to handle systematic missingness in distributed databases networks (DDNs, e.g., where a central analysis is not possible), by using methods such as MLMI.^{12,19} Secret et al.¹⁷ adapted this method and applied it in a DDN with simulated and real-world data using the UK's Clinical Practice Research Datalink. The authors concluded that this method can be used to reduce bias from

systematically missing data by allowing for statistical adjustment if at least one database contains the variables of interest. MLMI and other advanced MI methods have been shown to handle missing data and repair bias in a DDN, however, further research is required for its application in real pharmacoepidemiologic analysis.¹⁶⁻¹⁹

We believe that researchers, database management teams, reviewers and journal editors should be extra vigilant for data completeness. Compared to single-database studies, MDDBS have the additional complexity of heterogeneity between databases. To give insight into the existence of missing data, we recommend that researchers report the quantity and type of missing data for any variables that are relevant to the analysis or interpretation of results, in line with recommendations in the RECORD-PE and STROBE statements.^{20,21} In addition, missingness terms (MCAR, MAR, MNAR) should be used to simply demonstrate the potential impact of missing data on the study estimate. MDDBS should specifically report whether data are sporadically or systematically missing. This coupled with the use of an appropriate method for addressing missing data will increase overall confidence in the study.^{8,21,35} An overview of the recommended steps in addressing and reporting missing data in MDDB pharmacoepidemiologic studies can be found in Figure 2.

In this review, we successfully captured MDDB pharmacoepidemiologic studies from multiple regions around the world, expanding on previous work to identify these studies in a systematic search.¹ To our knowledge, a review of the reporting of missing data and the methods used to address it in MDDBS, has not previously taken place. However, there are some potential limitations to our study. First, we were not able to determine with absolute certainty how much data were missing in each study or database, as we were limited to reviewing only what was reported in published studies. Future research could assess the origin and quantity of missing data by directly examining pharmacoepidemiologic databases, particularly before and after data processing, and compare the findings against what is routinely reported for these databases. Second, we might have missed relevant studies due to difficulties in detecting MDDBS from systematic searches, as also indicated in similar studies.¹ For example, studies which use an established database network might not refer to the use of multiple databases in their abstract but instead to the network name. To account for this, we included names of well-known database networks in our search strategy, which were identified in collaboration with an expert. In addition, we only included publications in English, which could limit the generalisability of our findings beyond Europe and North America.

Multi-database pharmacoepidemiologic studies are deemed to be essential for regulatory and clinical assessments of drug safety and effectiveness, thus we must increase confidence in the potential that these studies can bring.³⁶ Missing data are a persistent problem in EHR, and it is underreported in multi-database pharmacoepidemiologic studies. The quantity and type of missing data as well as the resulting potential bias, and justification of the method used to address it should be reported.

ACKNOWLEDGEMENT

We would like to thank Xiaofeng Zhou (Pfizer Inc., Chair-elect ISPE Database SIG) for advice on specific database networks to search for this review.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICS STATEMENT

No ethical approval was required for the study.

ORCID

Nicholas B. Hunt  <https://orcid.org/0000-0003-1677-6025>

Helga Gardarsdottir  <https://orcid.org/0000-0001-5623-9684>

Romin Pajouheshnia  <https://orcid.org/0000-0002-4208-3583>

REFERENCES

- Bazelier MT, Eriksson I, de Vries F, et al. Data management and data analysis techniques in pharmacoepidemiological studies using a pre-planned multi-database approach: a systematic literature review. *Pharmacoepidemiol Drug Saf.* 2015;24:897-905.
- Platt RW, Platt R, Brown JS, Henry DA, Klungel OH, Suissa S. How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias. *Pharmacoepidemiol Drug Saf.* 2019;29:3-7.
- Pacurariu A, Plueschke K, McGettigan P, et al. Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. *BMJ Open.* 2018;8:e023090.
- Trifirò G, Coloma PM, Rijnbeek PR, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *J Intern Med.* 2014;275:551-561.
- Lai EC-C, Stang P, Yang YHK, Kubota K, Wong IC, Setoguchi S. International multi-database pharmacoepidemiology: potentials and pitfalls. *Curr Epidemiol Rep.* 2015;2:229-238.
- Toh S. Analytic and data sharing options in real-world multidatabase studies of comparative effectiveness and safety of medical products. *Clin Pharmacol Ther.* 2020;107:834-842.
- Klungel OH, Kurz X, De Groot MC, et al. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. *Pharmacoepidemiol Drug Saf.* 2016;25:156-165.
- Martín-Merino E, Calderón-Larrañaga A, Hawley S, et al. The impact of different strategies to handle missing data on both precision and bias in a drug safety study: a multidatabase multinational population-based cohort study. *Clin Epidemiol.* 2018;10:643-654.
- Perkins NJ, Cole SR, Harel O, et al. Principled approaches to missing data in epidemiologic studies. *Am J Epidemiol.* 2018;187:568-575.
- Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR Med Informatics.* 2018;6:e11.
- Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in using electronic health record data for CER. *Med Care.* 2013;51:S80-S86.
- Resche-Rigon M, White IR, Bartlett JW, Peters SAE, Thompson SG. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Stat Med.* 2013;32:4890-4905.
- Requena G, Abbing-Karahagopian V, Huerta C, et al. Incidence rates and trends of hip/femur fractures in five European countries: comparison using E-healthcare records databases. *Calcif Tissue Int.* 2014;94:580-589. <https://doi.org/10.1007/s00223-014-9850-y>.
- Khare R, Utidjian L, Ruth BJ, et al. A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Inform Assoc.* 2017;24:1072-1079.
- Pedersen A, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol.* 2017;9:157-166.

16. Jolani S. Hierarchical imputation of systematically and sporadically missing data: an approximate Bayesian approach using chained equations. *Biom J*. 2018;60:333-351.
17. Secrest MH, Platt RW, Reynier P, Dormuth CR, Benedetti A, Filion KB. Multiple imputation for systematically missing confounders within a distributed data drug safety network: a simulation study and real-world example. *Pharmacoepidemiol Drug Saf*. 2019;27:1-10.
18. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res*. 2018;27:1634-1649.
19. Audigier V, White IR, Jolani S, et al. Multiple imputation for multilevel data with continuous and binary variables. *Stat Sci*. 2018;33:160-183.
20. Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ*. 2018;363:1-22.
21. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med*. 2007;4:e296.
22. EMA. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCEPP). <http://www.encepp.eu/>. Accessed March 27, 2020.
23. U.S. Food & Drug Administration. Sentinel Initiative. <https://www.sentinelinitiative.org/>. Accessed March 27, 2020.
24. HealthCanada. Canadian Network for Observational Drug Effects Studies (CNODES). <https://www.cnodes.ca/>. Accessed March 27, 2020.
25. Asian Pharmacoepidemiology Network (AsPEN). <http://www.aspenet.asia/>. Accessed March 27, 2020.
26. The National Patient-Centered Clinical Research Network (PCORnet). <https://pcornet.org/>. Accessed March 27, 2020.
27. CDC. Vaccine Safety Datalink (VSD). <https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/vsd/>. Accessed March 27, 2020.
28. Gini R, Sturkenboom MC, Sultana J, et al. Different strategies to execute multi-database studies for medicines surveillance in real world setting: a reflection on the European model. *Clin Pharmacol Ther*. 2020;108:228-235. <https://doi.org/10.1002/cpt.1833>.
29. Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581-592.
30. Rioux C, Lewin A, Odejimi OA, Little TD. Reflection on modern methods: planned missing data designs for epidemiological research. *Int J Epidemiol*. 2020;49:1702-1711. <https://doi.org/10.1093/ije/dyaa042>.
31. Eekhout I, de Boer RM, Twisk JWR, de Vet HCW, Heymans MW. Missing data: a systematic review of how they are reported and handled. *Epidemiology*. 2012;23:729-732.
32. Spence AD, Busby J, Hughes CM, Johnston BT, Coleman HG, Cardwell CR. Statin use and survival in patients with gastric cancer in two independent population-based cohorts. *Pharmacoepidemiol Drug Saf*. 2019;28:460-470.
33. Dydensborg Sander S, Nybo Andersen AM, Murray JA, Karlstad Ø, Husby S, Størdal K. Association between antibiotics in the first year of life and celiac disease. *Gastroenterology*. 2019;156:2217-2229.
34. Kingwell E, Leray E, Zhu F, et al. Multiple sclerosis: effect of beta interferon treatment on survival. *Brain*. 2019;142:1324-1333.
35. Benchimol EI, Smeeth L, Guttman A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med*. 2015;12:e1001885.
36. European Medicines Agency. Post-authorisation safety studies (PASS). <https://www.ema.europa.eu/en/human-regulatory/post-authorisation/pharmacovigilance/post-authorisation-safety-studies-pass-0>. Accessed June 2, 2020.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Hunt NB, Gardarsdottir H, Bazelier MT, Klungel OH, Pajouheshnia R. A systematic review of how missing data are handled and reported in multi-database pharmacoepidemiologic studies. *Pharmacoepidemiol Drug Saf*. 2021;30:819-826. <https://doi.org/10.1002/pds.5245>