Clinical Epidemiology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Mind the Gaps: Literature Survey Reveals Shortcomings in Handling Missing Data in Clinical Practice Research Datalink (CPRD), a UK Primary Care Health Records Database

Esther Tolani[1], Miriam Rachel Carpenter[1], Irene Petersen[2,3], James R Carpenter[1,4]

[1]Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK; [2]Research Department of Primary Care and Population Health, University College London, London, UK; [3]Department of Epidemiology, Aarhus University, Aarhus, Denmark; [4]MRC Clinical Trials Unit, University College London, London, UK

Correspondence: Esther Tolani, Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK, Email esther.tolani@lshtm.ac.uk

Background: In the United Kingdom (UK) primary care electronic health records (EHR), key demographic, clinical, and lifestyle variables such as ethnicity, social deprivation, body mass index and smoking status are often incomplete. This incompleteness can compromise research validity by introducing bias and reducing statistical power. There are a number of frequently used approaches to handling missing data, including complete records analysis (CRA), missing indicator method and multiple imputation (MI), however it is not clear to what extent these are used in primary care EHR analyses or whether their use is appropriate. This study examines current practice for applying methodologies and reporting of missing data, in one of the largest UK primary care EHR databases, the Clinical Practice Research Datalink (CPRD).

Methods: A random ~10% sample of observational studies from the CPRD bibliography, published between 01 January 2013 and 31 December 2023, was selected. Article screening and data extraction for each paper was completed by two reviewers, who used pre-prepared pro-forma to independently extract reporting and methods for handling missing data.

Results: From 2,481 publications during the study period, a random 220 were selected for detailed review. Missing data were reported in 163 (74%) studies. CRA was applied in 50 studies (23%), missing indicator method was used in 44 studies (20%), MI in 18 studies (8%), and alternative methods such as reclassification and mean imputation, in 15 studies (6%).

Conclusion: Many studies fail to follow published best practice, often relying on flawed methods like the missing indicator method. Greater transparency, rigorous missing data techniques, and clearer reporting are needed. Improved guidance with practical examples would enhance research quality. Without methodological consistency and scrutiny, the risk of bias and misinterpretation remains high, making it essential to integrate missing data considerations into study design and analysis.

Keywords: complete records analysis, multiple imputation, observational research studies, primary care

Introduction

Observational data are a key resource for health researchers investigating possible relationships between exposures and outcomes. Routinely collected electronic health records (EHRs), such as the Clinical Practice Research Datalink (CPRD), a primary care EHR database, are widely used for research in the United Kingdom (UK).[1] Despite their widespread use, missing data remain a persistent challenge, in both CPRD and similar large administrative databases.[2] Alongside loss of precision and statistical power, missing data can lead to biased inferences, particularly when related to outcome, exposure and/or confounders.[3] This issue is especially relevant in studies of diseases where prevalence and exposure effects vary across key demographic, clinical, and lifestyle factors.[4] Given that sociodemographic factors like ethnicity, index of multiple deprivation, clinical and lifestyle factors such as body mass index (BMI), smoking, blood pressure are often incompletely recorded,[5–7] this raises concerns about the validity and generalisability of findings.

Received: 8 April 2025
Accepted: 4 October 2025
Published: 4 November 2025

There are several analytical approaches to analyse partially observed data. These make varying assumptions about the missingness mechanism at play in the data source and how it acts upon the outcome, exposure and confounders in the scientific model.[8] Rubin[9] first classified missingness mechanisms as missing not at random (MNAR), missing at random (MAR) and missing completely at random (MCAR), and in complex data sets such as EHRs, different mechanisms may cause the missingness of different variables. Under an MCAR mechanism, missingness is caused by factors which are independent of those under investigation (and hence independent of variables included in statistical models to address these questions). Under an MAR mechanism, the probability that data are missing may depend on the observed values but, crucially, this dependence is fully explained once we control for those variables. Informally speaking, any systematic differences can be explained by associations with the observed data. When the missingness mechanism is MNAR, given the fully observed data, the probability that data are missing remains dependent on the unobserved values of the incomplete variable. The challenge is that it is not possible to evaluate if the missing data are MAR or MNAR using only the available data.[10–12]

Given this, addressing missing data, by clearly stating assumptions, using appropriate methods and transparently reporting the results is not an optional extra, but a fundamental requirement for ensuring the reliability of epidemiological and clinical research. The consequences of failing to do so are not merely theoretical but have been demonstrated in practice. Poorly addressed missing data can lead to misleading inferences, underscoring the need for transparent reporting of assumptions and methods. For example, the initial QRISK study, developed to predict cardiovascular diseases revealed substantial missingness in key variables. Although the authors did multiple imputation, they failed to specify the multiple imputation model correctly.[13] This led to the authors erroneously reporting that serum cholesterol ratio was not an independent predictor of cardiovascular risk, highlighting how incomplete or inconsistently recorded data can undermine the reliability of clinical decision-making tools.[14] Furthermore, the robustness of the results to a range of contextually plausible assumptions about the missing data should be explored.

Over the years, numerous methods have been developed to improve handling missing data in research studies, and from at least 2009 there have been a number of guidance papers.[3,15–17] One of the key guidelines, the Treatment And Reporting of Missing data in Observational Studies (TARMOS) framework,[3] provides a structured approach for minimising the impact of missingness on study validity. This framework emphasises three key stages: planning, conducting and reporting the analysis:

1. Planning the analysis: This stage involves defining research questions, outlining statistical models, and pre-specifying methods to address missing data. Researchers are encouraged to consider potential missing data mechanisms (eg, MCAR, MAR, MNAR) and select appropriate analysis strategies.
2. Conducting the analysis: During this stage, researchers explore patterns of missing data and apply preplanned methods to address missing data. Sensitivity analyses are recommended to test the robustness of conclusions under different missing data assumptions.
3. Reporting the analysis: Transparent documentation of missing data handling is essential. This includes detailing imputation techniques, reporting the proportion of missing data, and evaluating how different methods impact study outcomes.

Key components highlighted by other frameworks include exploring the missingness patterns and performing a sensitivity analysis to explore the robustness of the conclusions.[18,19]

Among the various statistical methods available, multiple imputation (MI) is widely regarded as a robust, flexible and practical approach.[20] The most common approach and default in statistical software, is complete records analysis (CRA), which excludes individuals with missing values on one or more variables from the analysis. CRA is a natural starting point, but only valid under quite restrictive assumptions, and therefore typically insufficient. Alternatively, the missing data indicator method, primarily used for categorical variables, adds an additional category (eg, "value not observed" or "missing") to the categorical variable at hand. While this method allows researchers to retain all individuals in the analysis, the resulting inferences are generally inaccurate.[19] A third common method is single value imputation. This approach replaces missing values by a single common value: for example, by the observed mean for a continuous variable, or by the

most common category for a categorical variable. As an example of the latter, an individual whose ethnicity is missing may be imputed as "white".[21] Another method used is reclassification which involves reassigning existing values to different categories based on a schema or rule for example merging values into broad groups such as "white" and "non-white". Suboptimal approaches to handle missing data include single imputation, last observation carried forward and replacing missing values with "best" or "worst" values.[8] Lastly, inverse probability weighting (IPW) can be used to correct for bias due to missing data while preserving statistical integrity,[22] but in many applications it gives less precise results than multiple imputation.[12] This is because standard IPW (i) only retains complete records, reweighted to represent the full weighted sample and (ii) partially observed variables cannot be readily included in the weights.

While effective handling of missing data is crucial for ensuring the credibility and reliability of research findings, previous evidence suggests that current practice often falls short.[3] For instance, Graham[23] underscores the importance of transparent reporting and adherence to best practices in missing data analysis to maintain the robustness and validity of research findings. Despite the availability of these frameworks, many studies fail to fully adhere to the recommended guidelines, raising concerns about the accuracy of their results.[11,24] The widespread variability reported in missing data handling raises concerns about the appropriateness of current practices in EHR research.

More recent reviews have examined specific aspects of management of missing data in observational studies. Mainzer et al[25] conducted a scoping review of MI usage in causal inference studies, identifying substantial gaps in the reporting of imputation model specifications. Similarly, Wu et al[26] reviewed missing data reporting in UK critical care cohort studies, highlighting persistent deficiencies in transparency and methodological rigor. In the same vein, Okpara et al[27] reports a methodological survey of geriatric journals and found that 62.5% of these studies offered no clear statement or transparent reporting of missing data issues. In pharmacoepidemiologic studies using EHR Hunt et al,[28] reviewed 62 papers to assess reporting practices and statistical approaches for handling missing data. The authors found that missing data were handled inadequately and inconsistently.

However, no recent study has systematically examined missing data management within CPRD, which is one of the most widely used electronic health records databases for health research, nor has there been a comprehensive evaluation of adherence to best-practice guidelines in this context.

CPRD covers approximately 24% of the UK population, with coverage primarily concentrated in England.[1] It is one of the key resources for healthcare research in the UK, containing anonymised, linked individual patient data from both NHS primary and secondary care settings. CPRD research has informed drug safety guidance and clinical practice and resulted in approximately 3,500 peer-reviewed publications between 1988 and 2023. Additionally, CPRD's linkage to other datasets, such as hospital admissions and mortality records from the Office for National Statistics, enhances its utility for conducting detailed and comprehensive analyses. Given the increasing reliance on CPRD for health research and policy development, understanding how missing data are managed in this context is essential to ensure the robustness of research findings.[25,26]

This study aims to fill this gap through a detailed review of a random sample of research articles utilising CPRD data and listed in the CPRD bibliography. These were published in a wide range of journals, and include a range of studies, although our focus was on cohort, cross-sectional, and case-control studies. We investigated the prevalence of missing data, the methodologies employed to handle it, and the extent to which studies adhere to established guidelines. By identifying dominant methodologies, assessing their appropriateness, and evaluating reporting quality, this review highlights gaps in current practice and offers recommendations for improving missing data management in UK primary care research. Ultimately, this study seeks to support the process of aligning research methodologies with best-practice standards, to avoid inappropriate handling of missing data leading to misleading scientific conclusions.

Methods

Sampling Frame

The source of scientific publications for this study is the CPRD bibliography, downloaded from https://www.cprd.com/bibliography on 03/04/2024. The bibliography includes all peer reviewed papers published using CPRD data from the inception of the database in 1987. This bibliography, maintained by CPRD, is updated monthly. The CPRD bibliography is therefore a comprehensive source of publications for this study.

Selection of papers was carried out in two phases. In the first phase, we restricted to publications between 01 January 2013 and 31 December 2023. This time frame was chosen to allow us to explore how practice has changed over time. After applying the time restriction, there were 2,481 papers. Since this was too many for detailed review, a random sample of 220 papers (~10%) was extracted (Figure 1).

Paper Selection

The paper selection process used the software *Rayyan*.[29] Rayyan provided an automated randomised process which was used to extract a proportion of the papers. Using pre-defined selection criteria (see below), from the random sample of studies, we selected those classified as case-control, cohort, or cross-sectional studies. To check consistency and accuracy in the application of the selection criteria, a randomly selected sample of 20 articles was independently screened for selection by three of the four reviewers (ET, JC and IP).

Inclusion Criteria

- Published between 01 January 2013 and 31 December 2023
- Case control, cohort and cross-sectional study design
- May include comparisons or linkage to other secondary database sources

Exclusion Criteria

- Enhance any missing data with primary data collection
- Genomic research studies, eg, genome-wide association studies
- Clinical trials studies (for example, informing the selection of suitable individuals to participate in a clinical trial or use of CPRD data for individuals recruited in clinical trial)
- Systematic reviews

Data Extraction

Having identified the sample of studies for inclusion, a pilot data extraction process was conducted on ten papers by all reviewers (ET, JC, IP, and MC) to develop the data collection pro forma and establish a consistent method for completion. Then double data extraction was independently carried out for all 220 papers by two reviewers (ET and MC) (see Table S1). After the first set of 20 papers had been reviewed, ET and MC met to check agreement, resolve disagreements, and clarify the questions to reduce future disagreement. At the end of the review process ET and MC met to resolve disagreements. A small number of remaining disagreements were resolved by discussion among all authors.

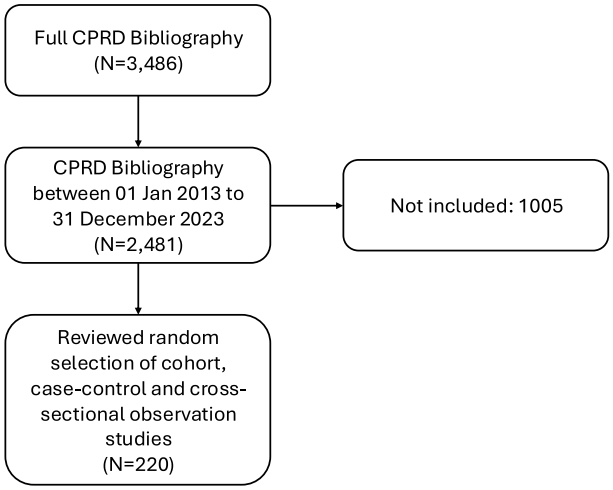

Figure 1 Study Review Process.

Extracted data were recorded and analysed using Microsoft Excel 24. The reporting process adheres to the guidelines outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist.

Information was grouped by theme: study metadata, study design, primary statistical analysis, handling of missing data in primary analysis, sensitivity analysis for missing data, study limitations, study variables (including confounders, exposures and outcome variables), and reporting transparency.

For each study, we recorded the total number of individuals who satisfied the study inclusion criteria as the study size, while we recorded the number of individuals included in the primary analysis as the analysis set size. Statistical analyses for the primary outcome were categorised into descriptive, regression (ie, generalised linear model, generalised linear mixed model, Cox proportional hazards model) analysis or other methods.

For missing data, we extracted the proportion of missing data in the study outcome, exposure and confounders, the method for handling missing data in the primary analysis (eg, multiple imputation or restricting to the subset of complete records), whether the assumptions about missing data underpinning the primary analysis were described/discussed and whether sensitivity analyses were performed; to explore the robustness of conclusions to a range of assumptions about the missing data. The full data collection pro forma is given in Table S2.

Results

Study sizes ranged between 76 and 17,480,766 individuals. Most papers were cohort studies (177, 81%), while the remaining studies consisted of case-control (35, 16%) and cross-sectional designs (8, 4%). The most frequent analysis was regression analysis (171 studies, 78%). Missing data was a prevalent issue, reported in 164 (75%) of studies. Despite this, explicit strategies for handling missing data were often absent, with 53 studies (24%) not providing any discussions of methods for handling missing data. For the primary analysis, MI was applied in 18 studies (8%), while 44 studies (20%) used a missing category indicator, and 50 studies (23%) conducted CRA. No relationship was observed between missing data method and study size (Table S3). Sensitivity analyses to assess the impact of missing data were performed in 24 studies (11%). Reporting practices for missing data varied widely across the studies, revealing significant inconsistencies. While 125 studies (57%) presented missing data in a table, only 90 studies (41%) described missing data or the associated analysis comprehensively in both text and table (Table 1).

Table 1 Study Characteristics

Characteristics		Number of Studies (%)
Study Characteristics	Study design	
	Case-control study	35 (15.9)
	Cross-sectional study	8 (3.6)
	Cohort study	177 (80.5)
	Analysis type	
	Regression analysis (% of regression analyses)	
	Survival analysis	89 (52.0)
	Logistic regression	47 (27.5)
	Linear regression	6 (3.5)
	Poisson regression	17 (9.9)
	Negative binomial regression	5 (2.9)
	Other	7 (4.10)
	Total (% of total studies)	171 (77.8)
	Descriptive analysis	31 (14.1)
	Other	18 (8.2)

(*Continued*)

https://doi.org/10.2147/CLEP.S528401

Table 1 (Continued).

Characteristics		Number of Studies (%)
	Study size	
	Min-Max	76 – 17480766
	Mean	525028
	Median	70622
	IQR	210282
	Not reported	12
	Analysis cohort size	
	Min-Max	76 – 17480766
	Mean	494324
	Median	69440
	IQR	196740
	Not reported	5
	Studies with difference between number of eligible individuals and analysis cohort size	22 (10)
Reporting Characteristics	Report missing data	163 (74.1)
	Report missing data in text only	36 (16.4)
	Report missing data in table only	35 (15.9)
	Report missing data in text and table	90 (40.9)

Study Characteristics

The 220 papers were published in 131 different journals. Most appeared in general medicine journals (60 studies, 27%), followed by specialist journals endocrinology (20 studies, 9%) and cardiology (15 studies, 7%). In terms of individual titles, the most common were the BMJ Open (14 studies, 6%) and British Journal of General Practice (10 studies, 5%) (see Table S4 for number of papers by journal specialty).

Studies were published between 2013 and 2023, with the highest proportions from 2019 (27 studies, 12%), and the fewest from 2013 (12 studies, 6%) (Figure 2). The study sizes ranged between 76 and 17,480,766 individuals, with a mean of 525,028 individuals, whilst mean analysis set size was 494,324 individuals. Twelve studies did not report the study size, and of these 5 studies did not report the analysis set size, so it was not possible to infer how many individuals were included in the analysis (Table 2).

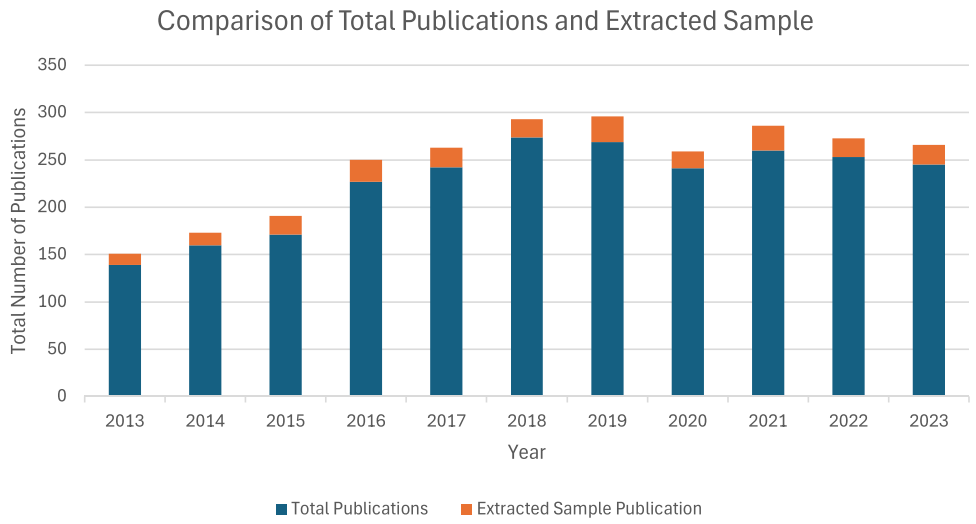

Figure 2 Publications in the Sample by Year.

Table 2 Method for Handling Missing Data by Study Type (Percentages of Studies Within Each Analysis Type is Given Within Each Row)

Study Method for Handling Missing Data in Primary Analysis	Descriptive Analysis (N=31)	Regression Analysis (N=171)	Other Methods (N=18)	Total Number of Studies (N=220)
Complete record analysis (N,%)	1 (2.0)	47 (94.0)	2 (4.0)	50 (100.0)
Missing indicator method (N,%)	2 (4.5)	38 (86.4)	4 (9.1)	44 (100.0)
Multiple imputation (N,%)	0 (0.0)	18 (100.0)	0 (0.0)	18 (100.0)
Other missing data method (incl. mean imputation, median imputation, reclassifying missing data) (N,%)	0 (0.0)	11 (73.3)	4 (26.7)	15 (100.0)
No method discussed or applied (N,%)	28 (30.1)	57 (61.3)	8 (8.6)	93 (100.0)

Regression analyses were utilised in 171 studies (78%), of these, the most common analysis was survival analysis (89 studies, 40.5% of all regression analyses), followed by logistic regression analysis (47 studies, 28% of all regression analyses). Descriptive analyses, which did not use statistical modelling, made up 31 (14%) of studies. Other types of analyses such as predictive modelling, post authorisation safety study and cost effectiveness studies were less common (18 studies, 8%).

Methodology for Handling Missing Data

Among the 163 studies (74%) that acknowledged missing data in one or more variables (Table 1), the most commonly used approach was the missing indicator method, applied in 44 studies (27%). CRA was used in 50 studies (30%), while MI was applied in the primary analysis of 18 studies (11%). Additionally, 15 studies (9%) employed other methods, such as reclassification and mean imputation (Table 2).

In 20 out of the 220 (9%) studies, there was a discrepancy between the number of eligible individuals and those included in the primary analysis, suggesting that CRA may have been used without this being explicitly stated (Table 1). Fifty-one (31%) studies stated how missing data informed the primary analysis (Figure 3). Sensitivity analyses addressing the impact of different assumptions about the missing data were performed in 25 out of the 220 (11%) studies, eg, by excluding variables from the model or using MI.

Of the studies reporting missing data (163 studies) there were 131 (80%) cohort studies, 24 (15%) case-control studies and 8 (5%) cross-sectional studies. MI was used in 16 out of these cohort studies (12%) and 2 (8%) out of these case-control studies. The missing indicator method was used in cohort 32 studies (24%), 10 case-control studies (42%) and 2 cross-sectional studies that reported missing data, respectively.

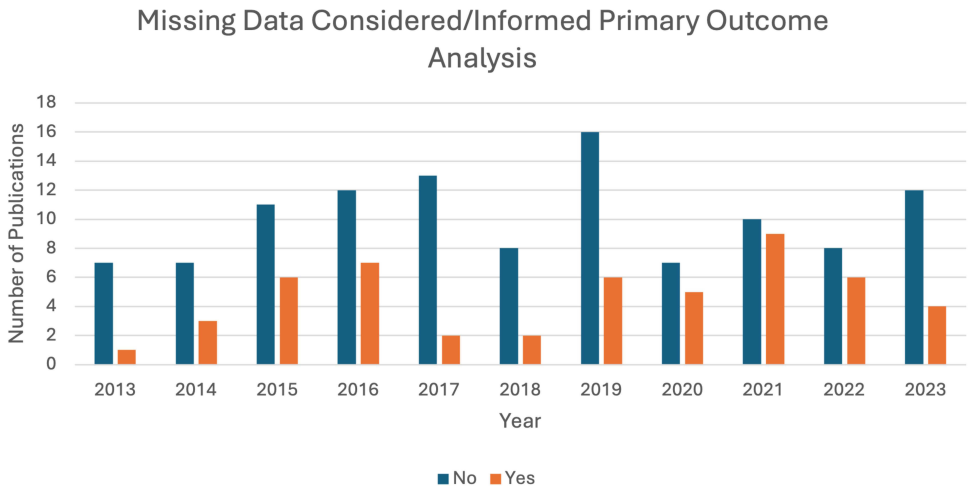

Figure 3 Missing Data Informed Primary Outcome Analysis.

https://doi.org/10.2147/CLEP.S528401

Other methods such as reclassification and mean imputation accounted for a small portion across all study designs (ie, cohort, case-control and cross-sectional), showing minimal variation in analysis approaches beyond those mentioned. In terms of missing data method used in different statistical analyses, for 2 (1%) descriptive analyses the missing indicator methods was used, followed by CRA in 5 (2%) studies, no methods for handling missing data were discussed for the remainder of the descriptive studies (Table 3). In terms of the most common regression method, survival analyses, the most common methods for handling missing data were missing indicator method (20 studies, 9%) and MI (14 studies, 6%).

Missing Data in Possible Confounding Variables

A total of 131 (60%) of the studies had missing data in possible confounding variables of the association between exposures and outcome, such as sociodemographic factors and key health indicators. Handling of missing data among possible confounding variables had the most diverse range of methods, suggesting greater variability in handling missing data for these variables. MI was more commonly applied in studies with missing confounders (15 studies). Missing data in outcome variables were relatively uncommon, with only 9 (4%) studies reporting missing outcome data (Table 4).

The health indicator, with the most missingness across studies was BMI which had a missing range of 1 to 90% (median 17% missing) across 98 (44% of all included) studies, with methods such as MI (18% of studies reporting BMI), addition of missing categories (29% of studies reporting BMI), and exclusion patients (11% of studies reporting BMI) being used to handle the missingness (Table 5). In terms of sociodemographic characteristics, among the 30 studies

Table 3 Missing Data Method by Study Designs (Percentages of Studies Within Each Method is Given Within Each Row)

Primary Analysis Missing Data Methodology	Study Type (% of Row)			
	Case-Control Study	Cohort Study	Cross-Sectional Study	Total
Multiple imputation	2 (11.1)	16 (88.8)	0 (0.0)	18 (100.0)
Missing indicator method	10 (22.7)	32 (72.7)	2 (4.5)	44 (100.0)
Complete record analysis	3 (6.0)	44 (88.0)	3 (6.0)	50 (100.0)
None/Not discussed	20 (21.5)	70 (75.3)	3 (3.2)	93 (100.0)
Other	0 (0.0)	15 (100.0)	0 (0.0)	15 (100.0)
Total	35 (15.9)	177 (80.5)	8 (3.6)	220 (100.0)

Table 4 Method for Handling Exposure, Outcome and Confounding Variables Across Studies

Variable Type	Method for Handling Missing Data (N, %)					
	Multiple Imputation	Missing Indicator Method	Reclassification	Complete Record Analysis	None/Not Discussed	Other
Exposure	2 (0.9)	3 (1.4)	1 (0.5)	10 (4.5)	8 (3.6)	2 (0.9)
Outcomes	0 (0.0)	1 (0.5)	1 (0.5)	2 (0.9)	5 (2.3)	0 (0.0)
Confounders	15 (6.8)	35 (15.9)	1 (0.5)	37 (16.8)	35 (15.9)	8 (3.6)

Note: Individual studies may report multiple types of missing variables; therefore, the total does not equal the number of studies.

Table 5 Missingness in Key Health Indicators and Sociodemographic Factors

Variable	Percentage Missing (Min-Max)	Methods for Handling	Number of Studies (%)
BMI	0.6–90.0	Total	98 (44.5)
		Complete record analysis	11 (5.0)
		Multiple imputation	19 (8.6)
		Missing indicator method	28 (12.7)
		None/Not discussed	32 (14.5)
		Other	8 (3.7)

(*Continued*)

Table 5 (Continued).

Variable	Percentage Missing (Min-Max)	Methods for Handling	Number of Studies (%)
Smoking	0.1–90.0	Total	95 (43.2)
		Complete record analysis	11 (5.0)
		Multiple imputation	15 (6.8)
		Missing indicator method	34 (15.5)
		None/Not discussed	30 (13.6)
		Other	5 (2.3)
Alcohol Consumption	0.6–85.6	Total	38 (17.3)
		Complete record analysis	4 (1.8)
		Multiple imputation	9 (4.1)
		Missing indicator method	15 (6.8)
		None/Not discussed	2 (0.9)
		Other	8 (3.6)
Ethnicity/Ethnic Group	1.4–82.0	Total	29 (13.2)
		Complete record analysis	2 (0.9)
		Multiple imputation	7 (3.2)
		Missing indicator method	10 (4.5)
		None/Not discussed	7 (3.2)
		Other	3 (1.4)
Index of Multiple Deprivation (IMD)	0.0–45.2	Total	17 (7.7)
		Complete record analysis	1 (0.5)
		Multiple imputation	4 (1.8)
		Missing indicator method	5 (2.3)
		None/Not discussed	7 (3.2)
		Other	0 (0.0)

reporting ethnicity data, the proportion of missing ethnicity ranged between 1% and 82% missingness (median 42%). The most frequent missing data methods used were adding a missing category (10, 33% of studies reporting ethnicity) or MI (7, 23% of studies reporting ethnicity). One study reclassified ethnicity into two broad groups: "white" and "non-white"[30] and another reclassified all missing ethnicity as "white".[31] Lastly, the Index of Multiple Deprivation (IMD) data was missing up to 45% of individuals (median 23%) across 17 (6% of all included) studies, with missing data handled through MI (4, 24% of studies reporting IMD) or the addition of a missing category (5, 29% of studies reporting IMD) (for more details on other key health indicators refer to Table 5).

Discussion

Strengths and Shortcomings of Current Practice

The results show both the variety of methods used to handle missing data in CPRD analyses and wide variation in the quality of reporting. Four studies[32–35] stood out because they followed guidelines by discussing the assumptions underpinning their primary CRA analysis. These articles provide useful examples for researchers to follow. By critiquing the assumptions their primary analyses rest on, and transparently reporting the methods used, they enhance confidence in the validity of their research findings.

Three studies[36–38] demonstrated good practice in their handling of missing data by providing sufficient information for their results to be reproduced. Consistent with key guidance papers[15] these studies outlined the MI procedure used, including the number of imputed datasets and the application of Rubin's rules. This strengthens research validity (c.f).[39,40]

One study stood out for reporting the extent of missing data and discussing potential missingness mechanisms.[41] Another study reported missingness for all variables, conducted MI, and specified the number of imputation datasets created. It also compared results before and after imputation, allowing the readers to see the robustness of the conclusions to different assumptions about the distribution of the missing data.[42]

Unfortunately, alongside these examples of good practice, around 80% of studies did not follow one or more key recommendations of the TARMOS framework for handling missing data. For example, a paper presented CRA as a sensitivity analysis, without explaining how the CRA assumptions differed from those used to justify the primary analysis.[43] While CRA is valid when the probability of a complete record, given the covariates, is independent of the response, the contextual plausibility of this assumption needs to be discussed. Further, because CRA are often very inefficient (limiting conclusions that can be drawn), it is usually useful to complement them with more efficient analyses, eg, multiple imputation.

One study reclassified missing ethnicity data into a separate category without assessing the potential bias introduced.[44] It also excluded data on IMD, often strongly associated with ethnicity, without further explanation or consideration of the implications. This mishandling of missing data raises concerns about the robustness of the findings.

Median/mean imputation, was another method used in four studies.[32,45–47] Because the median value of the variable is unrelated to other data from a patient, this approach biases the result, in ways that are not always easy to predict. For example, median imputation of missing values for confounder can mean that important unadjusted confounding remains, thus biasing the reported effects of exposure.[8]

The failure to provide justifications for excluding data was an issue in at least 20% of studies using complete case analysis (Table 3). For example, a study by Kostanjsek et al[48] which investigated whether undergoing bariatric surgery is associated with a reduced risk of developing new-onset heart failure among individuals with obesity excluded participants with missing BMI, a critical indication for bariatric surgery, without explaining the rationale. The question is whether this exclusion might have introduced a potential bias in their findings.

Within EHR settings, outcome variables which are the consequence of health care needs, ie, driven by disease incidence or recurrence are generally well captured. This is because they reflect underlying costs, and capturing these costs is the primary purpose of the database. Typically, when outcomes are not captured, they are deemed not to have occurred. However, this should not be uncritically assumed; it is important to remember that databases like CPRD only captures what is in the patient's electronic health records. Often, we find that incidence and prevalence of a number of conditions are slightly lower in electronic health records compared to community/population surveys.[49]

A similar issue arises with exposure variables, if an exposure is the result of a cost, it is likely to be recorded. For example, we can be relatively confident about drugs a patient has been prescribed (though not, of course, whether they took them). In the same way, we have also seen that people with chronic conditions (eg, diabetes, cardiovascular diseases, COPD, etc) covered by the Quality Outcome Framework (QOF) are more likely to have health indicators recorded on regular basis.[5] However, as with outcomes, studies typically make the implicit assumption that the reason for any missing exposure data is unrelated to the study question, so does not bias the results. This point is particularly relevant for conditions classified using syndromic definitions, where diagnosis is often inferred from a combination of recorded symptoms and prescriptions. If key components are not documented, the condition may be under- or misclassified rather than truly absent. Again, in our survey, we did not find any papers that discussed this point.

To mitigate the limitations in capturing outcomes and exposures, a natural approach is to seek supplementary information through data linkage of CPRD with complementary sources such as the Hospital Episode Statistics, or ONS records. In addition, when direct measurement is not available, population level imputation using linked external datasets can be used to improve multiple imputation of missing values.[21]

Across 170 studies (77%), the methodology for handling missing data was either poorly documented or entirely absent. For example, while 125 studies (57%) acknowledged missing data in tables only a few discussed it in their methods, results, or discussion sections, limiting reproducibility (e.g).[50] The lack of reporting also contributed to a recurring challenge during the review as it was unclear whether studies actually used CRA. Many papers mentioned missing data without clarifying how missing data were handled in subsequent analyses (e.g).,[51,52] which made it difficult to evaluate whether the methods employed appropriately addressed potential biases caused by missing data.

Adherence to Reporting Guidelines

Several guidelines have been developed to help authors with study design, methodology and analysis.[53,54] Reporting is covered by the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guideline and more

specifically an extension RECORD guideline was developed to provide a checklist for improving the transparency and completeness of reporting in studies using routinely collected health data. Although the guideline states that the authors should explain how missing data were addressed, they do not specifically require authors to articulate their assumptions regarding missing data.[55] This issue is particularly relevant for UK primary care studies, where additional guidelines and their consistent application are necessary.

Based on established best-practice guidelines for handling missing data,[56,57] we would have expected a more standardised and transparent approach across the reviewed studies. We had hoped to find that missing data reporting was typically following recommendations, with detailed documentation in tables and methodological sections outlining the proportion of missingness, missing data assumptions, and justification for and reproducible description of the methods used. Unfortunately, this was far from the case. Many studies fell short of meeting the reporting standards required for contemporary research practices for several reasons including not stating how missing data informed the primary outcome analysis (77%), or only reporting missing data in a table (16%). For instance, in the study by Kontopantelis et al,[58] it was difficult to determine the number of individuals included in the analysis, underscoring a lack of transparency in reporting key methodological details. Similarly, in studies where person-years were the primary outcome, it was often challenging to ascertain the number of individuals included in the denominator or the actual size of the cohort, further complicating the interpretation of the results (eg,).[59,60]

Ideally, researchers should pre-specify missing data strategies during study design, explicitly discussing the extent of missing data, whether data are plausibly MCAR, MAR, or MNAR, and how this informs the analysis. Given the frequent missingness observed in sociodemographic and key health indicator variables (which are key confounding covariates in many analyses), MI should have been the predominant method, as it is widely regarded as the most statistically valid approach when covariates are approximately missing at random, and a natural method for exploring sensitivity of conclusions to departures from this assumption, as recommended and illustrated in the TARMOS framework. Likewise, CRA should be limited to cases where its assumptions are plausible. The implementation of sensitivity analyses can be supported by the adoption of missingness-directed acyclic graphs (m-DAGs), which provide a structured and transparent approach to addressing uncertainty about missing data assumptions.[61]

One study combined the missing data with another category such as "other".[44] This practice introduces a loss of meaning by conflating unknown and infrequent categories. Further, since the new, combined, category contains a range of true (underlying) values, confounding may not be properly adjusted for. In other studies we were unable to determine any methods for handing missing data as the methods were unclear and the reporting was very ambiguous.[62,63]

Study Limitations

Given the study objectives, the only feasible approach was to carry out a detailed review of a random sample of papers. This also avoided the challenge of identifying papers for inclusion in the study by checking the abstracts and keywords for information about how missing data were handled. However, the challenges of missing data and their handling are likely to be similar across other EHR databases. For instance, Petersen et al demonstrated that health indicators are also frequently missing in another UK primary care database, The Health Improvement Network, and that the patterns of missing data may be associated with an individual's health status.[5] Nevertheless, future research could usefully confirm whether similar issues with missing data handling are present in publications using data from other EHR databases.

Another potential limitation of this study is the reliability of data extraction, as differences in interpretation could affect how missing data practices were classified. This concern was mitigated through a double-review process, where each paper was independently assessed by two researchers, and discrepancies were resolved through discussion (see methods). This approach helped enhance consistency and accuracy in data classification.

While a larger sample of papers would have allowed for a more extensive evaluation, the number of studies reviewed was sufficient to provide a reliable assessment of current missing data practices. The sample size ensured that the findings captured meaningful trends without introducing significant uncertainty, making it appropriate for addressing the study's objectives.

Conclusion

This study highlights the need for improvements in both the choice of analysis methods for, and reporting of, missing data CPRD studies. While it is true that not all papers aimed at a clinical audience are required to report on missing data, doing so remains a key aspect of methodological transparency. Missing data can introduce bias, reduce statistical power, and compromise the generalisability of findings. These issues are directly relevant to clinical interpretation and application. Given that clinical decisions and guidelines may be informed by such research, even brief reporting on the extent, handling, and potential impact of missing data strengthens the validity and utility of the findings. Encouraging consistent reporting of missing variables, particularly in studies using routine data, is therefore not just a methodological concern but a matter of clinical relevance. While some studies, eg, those highlighted above, are exemplary, it is disappointing that, despite guidelines for good practice being widely available, these are a small minority.

The widespread use of the missing indicator approach and the approach that assigns missing values of a categorical variable the most frequent value (eg, missing ethnicity assigned to "white") is particularly concerning, since it has been accepted in the literature for many years that these methods are not valid under plausible assumptions about the missing data and generally give misleading inferences.

In the light of this, we argue that a step-change is needed. Alongside dissemination of exemplars of good practice to guide researchers, we believe journal editors and reviewers have a key role to play. Three key points are (i) checking missing data are discussed in every manuscript; (ii) challenging authors who have used the missing indicator or most frequent category method; (iii) asking for justification of the assumptions for complete records analysis and/or multiple imputation. Ideally, in addition to these three points, authors, reviewers and editors should ask themselves if the conclusions are likely sensitive to the missing data assumptions, and if so, consider a sensitivity analysis.

With large observational datasets, missing data are inevitable; for ~30% of papers to ignore this issue is not acceptable Addressing missing data must be recognised as fundamental to producing high-quality, reliable research in primary care. Without greater methodological consistency, transparency, and scrutiny, the risk of bias and misinterpretation will persist. To drive meaningful change, researchers, journals, and institutions must work together to establish and enforce higher standards, ensuring that the handling of missing data is no longer an afterthought, but a core element of study design and reporting.

Abbreviations

CPRD, clinical practice research datalink; COPD, chronic obstructive pulonmary disease; UK, United Kingdom; EHR, electronic health records; CRA, complete records analysis; MI, multiple imputation; TARMOS, treatment and reporting of missing data in observational studies; MCAR, missing completely at random; MNAR, missing not at random; MAR, missing at random; PRISMA-ScR, preferred reporting items for systematic reviews and meta-analyses extension for scoping reviews; BMI, body mass index; IMD, index of multiple deprivation; IQR, interquartile range; STROBE, strengthening the reporting of observational studies in epidemiology.

Acknowledgments

ChatGPT (GPT-4)[63] was used to proofread and improve the language of the early manuscript drafts, this was iterated upon several times by the primary author (Esther Tolani).

This work was supported by the funding of IQVIA to Esther Tolani, none of the funding sources contributed to the study design, collection, analysis and interpretation of data, writing of the report and decision to submit the article for publication.

James Carpenter is funded by MRC grant MC_UU_00004/07.

Disclosure

Professor James Carpenter reports grants from UK Medical Research Council, personal fees from Wiley, personal fees from Springer, personal fees from University of Bern, personal fees from Statisticians in the Pharmaceutical Industry, personal fees from Novartis, during the conduct of the study. The author(s) report no conflicts of interest in this work.

References

1. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: clinical practice research datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827–836. doi:10.1093/ije/dyv098
2. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLOS Med*. 2015;12(10):e1001885. doi:10.1371/journal.pmed.1001885
3. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: the treatment and reporting of missing data in observational studies framework. *J Clin Epidemiol*. 2021;134:79–88. doi:10.1016/j.jclinepi.2021.01.008
4. Cooper J, Nirantharakumar K, Crowe F, et al. Prevalence and demographic variation of cardiovascular, renal, metabolic, and mental health conditions in 12 million English primary care records. *BMC Med Inf Decis Mak*. 2023;23(1):220. doi:10.1186/s12911-023-02296-z
5. Petersen I, Welch CA, Nazareth I, et al. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clin Epidemiol*. 2019;11:157–167. doi:10.2147/CLEP.S191437
6. Mathur R, Bhaskaran K, Chaturvedi N, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *J Public Health Oxf*. 2014;36(4):684–692. doi:10.1093/pubmed/fdt116
7. Shiekh SI, Harley M, Ghosh RE, et al. Completeness, agreement, and representativeness of ethnicity recording in the United Kingdom's Clinical Practice Research Datalink (CPRD) and linked Hospital Episode Statistics (HES). *Popul Health Metr*. 2023;21(1):3. doi:10.1186/s12963-023-00302-0
8. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*. 2017;9:157–166. doi:10.2147/CLEP.S129785
9. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. 1987. doi:10.1002/9780470316696
10. Curnow E, Carpenter JR, Heron JE, et al. Multiple imputation of missing data under missing at random: compatible imputation models are not sufficient to avoid bias if they are mis-specified. *J Clin Epidemiol*. 2023;160:100–109. doi:10.1016/j.jclinepi.2023.06.011
11. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377–399. doi:10.1002/sim.4067
12. Little RJA, Carpenter JR, Lee KJ. A comparison of three popular methods for handling missing data: complete-case analysis, inverse probability weighting, and multiple imputation. *Sociol Methods Res*. 2022;004912412211138. doi:10.1177/00491241221113873
13. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*. 2007;335(7611):136. doi:10.1136/bmj.39261.471806.55
14. Li Y, Sperrin M, Belmonte M, Pate A, Ashcroft DM, van Staa TP. Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Sci Rep*. 2019;9(1):11222. doi:10.1038/s41598-019-47712-5
15. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. doi:10.1136/bmj.b2393
16. Carpenter JR, Smuk M. Missing data: a statistical framework for practice. *Biom J*. 2021;63(5):915–947. doi:10.1002/bimj.202000196
17. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol*. 2019;48(4):1294–1304. doi:10.1093/ije/dyz032
18. Carpenter JR, Kenward MG. Sensitivity Analysis: MI Unleashed. In: *Multiple Imputation and Its Application*. 2013:229–268. doi:10.1002/9781119942283.ch10
19. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol*. 1995;142(12):1255–1264. doi:10.1093/oxfordjournals.aje.a117592
20. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1–67. doi:10.18637/jss.v045.i03
21. Pham TM, Carpenter JR, Morris TP, Wood AM, Petersen I. Population-calibrated multiple imputation for a binary/categorical covariate in categorical regression models. *Stat Med*. 2019;38(5):792–808. doi:10.1002/sim.8004
22. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*. 2013;22(3):278–295. doi:10.1177/0962280210395740
23. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol*. 2009;60:549–576. doi:10.1146/annurev.psych.58.110405.085530
24. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol*. 2017;17(1):162. doi:10.1186/s12874-017-0442-1
25. Mainzer RM, Moreno-Betancur M, Nguyen CD, Simpson JA, Carlin JB, Lee KJ. Gaps in the usage and reporting of multiple imputation for incomplete data: findings from a scoping review of observational studies addressing causal questions. *BMC Med Res Methodol*. 2024;24(1):193. doi:10.1186/s12874-024-02302-6
26. Wu TT, Smith LH, Vernooij LM, Patel E, Devlin JW. Data missingness reporting and use of methods to address it in critical care cohort studies. *Crit Care Explor*. 2023;5(11).
27. Okpara C, Edokwe C, Ioannidis G, Papaioannou A, Adachi JD, Thabane L. The reporting and handling of missing data in longitudinal studies of older adults is suboptimal: a methodological survey of geriatric journals. *BMC Med Res Methodol*. 2022;22(1):122. doi:10.1186/s12874-022-01605-w
28. 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(7):819–826. doi:10.1002/pds.5245
29. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210. doi:10.1186/s13643-016-0384-4
30. Sultan A, West J, Ban L, et al. Adverse pregnancy outcomes among women with inflammatory bowel disease: a population-based study from England. *Inflamm Bowel Dis*. 2016;22(7):1621–1630. doi:10.1097/mib.0000000000000802
31. Habte-Asres HH, Murrells T, Nitsch D, Wheeler DC, Forbes A. Glycaemic variability and progression of chronic kidney disease in people with diabetes and comorbid kidney disease: retrospective cohort study. *Diabetes Res Clin Pract*. 2022;193:110117. doi:10.1016/j.diabres.2022.110117
32. Ashdown HF, Smith M, McFadden E, Pavord ID, Butler CC, Bafadhel M. Blood eosinophils to guide inhaled maintenance therapy in a primary care COPD population. *ERJ Open Res*. 2022;8(1):00606–2021. doi:10.1183/23120541.00606-2021

33. Khunti K, hertz CL, Husemoen LLN, et al. Cardiovascular risk factors early in the course of treatment in people with type 2 diabetes without established cardiovascular disease: a population-based observational retrospective cohort study. *Diabet Med*. 2021;39:e14697. doi:10.1111/dme.14697
34. Leal J, Murphy J, Garriga C, et al. Costs of joint replacement in osteoarthritis: a study using the National Joint Registry and Clinical Practice Research Datalink datasets. *Arthritis Care Res Hoboken*. 2020. doi:10.1002/acr.24470
35. Douglas IJ, Bhaskaran K, Batterham RL, Smeeth L. The effectiveness of pharmaceutical interventions for obesity: weight loss with orlistat and sibutramine in a United Kingdom population-based cohort. *Br J Clin Pharmacol*. 2015;79(6):1020–1027. doi:10.1111/bcp.12578
36. Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open*. 2013;3(8):e003482. doi:10.1136/bmjopen-2013-003482
37. Brunetti VC, Reynier P, Azoulay L, et al. SGLT −2 inhibitors and the risk of hospitalization for community-acquired pneumonia: a population-based cohort study. *Pharmacoepidemiol Drug Saf*. 2021;30:740–748. doi:10.1002/pds.5192
38. Hawley S, Leal J, Delmestri A, et al. Anti-osteoporosis medication prescriptions and incidence of subsequent fracture among primary hip fracture patients in England and Wales: an interrupted time-series analysis. *J Bone Min Res*. 2016;31(11):2008–2015. doi:10.1002/jbmr.2882
39. Akyea RK, Vinogradova Y, Qureshi N, et al. Sex, Age, and Socioeconomic Differences in Nonfatal Stroke Incidence and Subsequent Major Adverse Outcomes. *Stroke*. 2021;52(2):396–405. doi:10.1161/strokeaha.120.031659
40. Woods LM, Rachet B, Morris M, Bhaskaran K, Coleman MP. Are socio-economic inequalities in breast cancer survival explained by peri-diagnostic factors? *BMC Cancer*. 2021;21(1):485. doi:10.1186/s12885-021-08087-x
41. Blagojevic-Bucknall M, Mallen C, Muller S, et al. The risk of gout among patients with sleep apnea: a matched cohort study. *Arthritis Rheumatol*. 2018;71:154–160. doi:10.1002/art.40662
42. Akyea RK, Doehner W, Iyen B, Weng SF, Qureshi N, Ntaios G. Obesity and long-term outcomes after incident stroke: a prospective population-based cohort study. *J Cachexia Sarcopenia Muscle*. 2021;12:2111–2121. doi:10.1002/jcsm.12818
43. Bromley SE, Matthews A, Smeeth L, Stanway S, Bhaskaran K. Risk of dementia among postmenopausal breast cancer survivors treated with aromatase inhibitors versus tamoxifen: a cohort study using primary care data from the UK. *J Cancer Surviv*. 2019;13(4):632–640. doi:10.1007/s11764-019-00782-w
44. Morton A, Simpson A, Humes D. Regional variations and deprivation are linked to poorer access to laparoscopic and robotic colorectal surgery: a national study in England. *Tech Coloproctol*. 2023;28(1):9. doi:10.1007/s10151-023-02874-3
45. Coates G, Clewes P, Lohan C, et al. Health economic impact of moderate-to-severe chronic pain associated with osteoarthritis in England: a retrospective analysis of linked primary and secondary care data. *BMJ Open*. 2023;13(7):e067545. doi:10.1136/bmjopen-2022-067545
46. Herrett E, Shah AD, Boggon R, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ*. 2013;346:f2350. doi:10.1136/bmj.f2350
47. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944
48. Kostanjsek L, Ardissino M, Moussa O, et al. Bariatric surgery and incident heart failure: a propensity score matched nationwide cohort study. *Int J Cardiol*. 2023;378:42–47. doi:10.1016/j.ijcard.2023.01.086
49. Aldridge R, Evans H, Yavlinsky A, et al. Estimating disease burden using national linked electronic health records: a study using an English population-based cohort. [version 2; peer review: 2 approved]. *Wellcome Open Res*. 2024;8(262):262. doi:10.12688/wellcomeopenres.19470.2
50. Renoux C, Vahey S, Dell'Aniello S, Boivin JF. Association of selective serotonin reuptake inhibitors with the risk for spontaneous intracranial hemorrhage. *JAMA Neurol*. 2017;74(2):173–180. doi:10.1001/jamaneurol.2016.4529
51. Gorton HC, Webb RT, Carr MJ, DelPozo-Banos M, John A, Ashcroft DM. Risk of unnatural mortality in people with epilepsy. *JAMA Neurol*. 2018;75(8):929–938. doi:10.1001/jamaneurol.2018.0333
52. Peeters PJ, Bazelier MT, Leufkens HG, et al. Insulin glargine use and breast cancer risk: associations with cumulative exposure. *Acta Oncol*. 2016;55(7):851–858. doi:10.3109/0284186x.2016.1155736
53. Wang SV, Pottegård A, Crown W, et al. HARmonized protocol template to enhance reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: a good practices report of a Joint ISPE/ISPOR Task Force. *Value Health*. 2022;25(10):1663–1672. doi:10.1016/j.jval.2022.09.001
54. Low GK, Subedi S, Omosumwen OF, et al. Development and validation of observational and qualitative study protocol reporting checklists for novice researchers (ObsQual checklist). *Eval Program Plann*. 2024;106:102468. doi:10.1016/j.evalprogplan.2024.102468
55. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. *JAMA Network Open*. 2021;4(2):e210184. doi:10.1001/jamanetworkopen.2021.0184
56. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335(7624):806–808. doi:10.1136/bmj.39335.541782.AD
57. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. *Lancet*. 2008;371(9619):1149–1150. doi:10.1016/s0140-6736(08)60505-x
58. Kontopantelis E, Springate D, Ashcroft D, et al. Associations between exemption and survival outcomes in the UK's primary care pay-for-performance programme: a retrospective cohort study. *BMJ Qual Saf*. 2016;25(9):657–670. doi:10.1136/bmjqs-2015-004602
59. O'Sullivan JW, Stevens S, Hobbs FDR, et al. Temporal trends in use of tests in {UK} primary care, 2000-15: retrospective analysis of 250 million tests. *BMJ*. 2018;363:k4666. doi:10.1136/bmj.k4666
60. Bietry FA, Hug B, Reich O, Susan JS, Meier CR. Iron supplementation in Switzerland - A bi-national, descriptive and observational study. *Swiss Med Wkly*. 2017;147:w14444. doi:10.4414/smw.2017.14444
61. Lee KJ, Carlin JB, Simpson JA, Moreno-Betancur M. Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification. *Int J Epidemiol*. 2023;52(4):1268–1275. doi:10.1093/ije/dyad008
62. Renoux C, Shin JY, Dell'Aniello S, Fergusson E, Suissa S. Prescribing trends of attention-deficit hyperactivity disorder (ADHD) medications in UK primary care, 1995-2015. *Br J Clin Pharmacol*. 2016;82(3):858–868. doi:10.1111/bcp.13000
63. OpenAI. GPT (GPT-4) [Large language model]. Retrieved from https://openai.com. Accessed 7 March, 2023.