Using routinely collected data for research purposes: challenges and mitigation strategies

Sabine Hoffmann,[1] Tim Morris,[2] Moritz Herrmann,[3,4] Georg Heinze,[5] Laure Wynants,[6,7,8] Ben Van Calster,[7,8] Bernd Bischl,[1,4] Matthias Schmid,[9] Pamela A Shaw,[10] Tim Mathes,[11,12] Florian Naudet,[13,14] Frank E Harrell,[15] Mona Niethammer,[1] Samuel Muli,[16] Sebastian Zimmer,[17] Farhad Bakhtiary,[18] David Leistner,[19] P Christian Schulze,[20] Daniel Sedding,[21] Philipp Lurz,[22] Tienush Rassaf,[23] Malte Kelm,[24] Stephan Baldus,[25] Johann Bauersachs,[26] Georg Nickenig,[17] Holger Thiele,[27] Enzo Lüsebrink[17]

For numbered affiliations see end of the article

Correspondence to: S Hoffmann sabine.hoffmann@stat.uni-muenchen.de (or @sabinelk11 on X; ORCID 0000-0001-6197-8801)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2026;393:e087812
http://dx.doi.org/10.1136/bmj-2025-087812

Accepted: 25 March 2026

Increasing availability of large routinely collected datasets presents many possibilities to answer more questions about health and disease, and at a faster pace. These opportunities are exciting, but, without the necessary expertise, well intentioned researchers can unwittingly fall into traps that make their work indistinguishable from that of less well meaning researchers. This article describes important challenges that arise in the analysis of routinely collected data and offers mitigation strategies to improve the trustworthiness of experimental and observational studies with the aim of explanation or prediction using classical statistical methods or AI algorithms. It also provides a roadmap to support researchers in conducting analyses on routinely collected data that produce more reliable estimates.

SUMMARY POINTS

Routinely collected data, which are increasingly used for research purposes, offer significant opportunities in terms of time and cost of data collection, but researchers often lack knowledge of how the data were generated and control over how they were collected

Common challenges in the analysis of routinely collected data include representativeness, time point alignment, data quality, interventions and tests not being random, and analytical variability due to a multiplicity of possible analysis strategies

The flexibility of AI algorithms can make them highly vulnerable to the biases arising in the analysis of routinely collected data

Without rigorous internal and external validation, and in the absence of clinical and methodological expertise, these AI tools may do more harm than good when analysing routinely collected datasets

Lack of data integrity may mean that prospectively collecting research data will often be better, as no evidence for a research question may be preferable to evidence from a low quality study that may be misleading and over-interpreted

The ever increasing availability of large datasets allows for exciting opportunities for research to enhance human health. Studies based on routinely collected data use datasets that are not collected for research purposes, such as electronic health records, administrative claims and billing data, registries, or information from wearables and apps. We use the term routinely collected data instead of the widely used terms real world data and real world evidence, because it avoids the well documented ambiguity surrounding these terms, for which numerous and inconsistent definitions exist.[1] In line with this, the term "real world" is being used for study designs as diverse as systematic reviews, case series, randomised controlled trials, and cross sectional diagnostic accuracy studies,[2] whereas major reporting guidelines recommend that the study design be explicitly identifiable.[3] To avoid conflating data source with study design, we therefore use the term routinely collected data, which more accurately refers to the origin of the data rather than implying a particular type of evidence. Routinely collected data have potential for each of three common tasks of data analysis[4]: description (what are the characteristics of a certain group and how are they related?); explanation (how does an intervention affect outcomes?); and prediction (how precisely can future events be forecasted or current states of health be diagnosed[5]?).

Routinely collected data raise hopes of so-called real world evidence that can optimise personalised treatment regimens at scale by recognising patterns and rare events, using long term outcomes. Advances in computing power and rapid developments in the field of AI fuel the allure of automated analyses that will place data driven, personalised medicine within reach.[6] Although AI and digitalization in health hold great potential for reducing costs, providing broader access to care and improving data quality in routinely collected data, enthusiasm about data and algorithms has been dampened by sobering warnings that, without expertise in study design, data collection, and data analysis, throwing algorithms at large quantities of data may do more harm than good.[7-9] Contradictory, overoptimistic, and unreplicable findings are to be expected, and health disparities may be exacerbated.[10] Childers and Maggard-Gibbons discuss the results of two teams of researchers who considered the same research question on the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) but used different outcome measures, inclusion and

exclusion criteria, covariates, and operationalisation of covariates and obtained opposite findings that were published in the same year and the same journal.[11] Several automated early warning systems that are being integrated in electronic health records have shown disappointing performance in practice.[12 13] Finally, analyses of routinely collected data can give rise to so-called big data paradoxes whereby "the more the data, the surer we fool ourselves"[14], producing seemingly very precise but biased results that are far from the true value being estimated. This has best been exemplified with studies for prediction of influenza-like illness and the estimation of vaccine uptake.[15 16] Overall, concerns exist that the availability of big routinely collected data in combination with powerful AI algorithms will lead to a deluge of scientific publications of dubious value, drowning out trustworthy research findings and undermining the credibility of health research.[17 18]

Researchers need tools to critically appraise studies using routinely collected data. Several reporting guidelines and checklists exist for the analysis of routinely collected data,[19-21] but little practical guidance is available on how to evaluate and to improve their quality in light of the challenges that arise. Moreover, previous efforts to discuss these challenges have largely focused on either experimental or observational studies with the aim of explanation or prediction using either classical statistical methods or AI algorithms. Considering these different designs, tasks, and methods separately may create false dichotomies. As they are often affected by the same problems, we argue that considering them together is more fruitful. This paper discusses important questions, suggests guiding principles for interpretation, and, where possible, describes strategies to overcome challenges and limitations. It also provides a roadmap to support researchers in conducting analyses on routinely collected data that produce more reliable estimates.

Challenges in the analysis of routinely collected data

> The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.
>
> John Tukey

Routinely collected data offer significant opportunities in terms of time and cost of data collection, but appropriate analysis and interpretation of results are challenging. When data are prospectively collected for research purposes, the design can be aligned with the aims of the study by enrolling participants that ensure a representative sample and by pre-specifying standardised procedures to measure all relevant variables. Routinely collected data are usually collected for a completely different purpose, such as administration or clinical documentation, and do not involve any additional prospective data collection. As a result, researchers have little control over the selection of patients included in the data, the selection of variables that are measured, and when and how they are measured. Treatments are typically not randomly assigned, reasons for treatment decisions may not be recorded, masking is typically not possible, and measurements, including those of outcomes, are usually not standardised but rather reflect results of everyday clinical decision making. Box 1 provides evidence and examples of challenges arising in the analysis of routinely collected data, and supplementary table A provides an overview of risks and mitigation strategies.

Representativeness and generalisability

Inclusion and exclusion criteria for studies using routinely collected data are usually dictated, or at least limited, by data availability. Despite the promise of routinely collected data to maximise representativeness and generalisability by providing evidence under "real world" conditions, these datasets are often collected in specific contexts (for example, university hospitals, statutory insured) and underrepresent or misrepresent certain groups of patients.[22 63] As shown in box 1, evidence shows disparities in the access to and the use of healthcare services. Moreover, substantial variation may exist in patients' characteristics and in disease occurrence across sites in multisite studies.[64 65] Even if the routinely collected data include a representative sample from the target population, identifying patients will often be challenging because the data simply lack information to establish inclusion and exclusion criteria,[66 67] potentially requiring changes in the research question that can be answered with the data.

When the aim is explanation, a dataset that is unrepresentative of the target population may undermine external validity and affect the generalisability of results. This is particularly true when the distribution of treatment effect modifiers differs between the available data and the target population.[68] For both explanation and prediction, a non-representative sample is highly problematic when the selection is made on the basis of a so-called collider that may, for instance, be influenced by both exposure and outcome.[69 70] As pointed out by Griffith and colleagues,[69] restricting analyses to patients who have been admitted to hospital may attenuate, inflate, or reverse the sign of the association between any variables that relate to hospital admissions. In prediction, models that are developed on routinely collected data with high disease incidence may show poor calibration by overestimating risks.[65] Moreover, predictive performance may be poor for groups that are underrepresented or misrepresented in the data,[71 72] potentially exacerbating existing health disparities.[10 73]

Tackling representativeness and generalisability

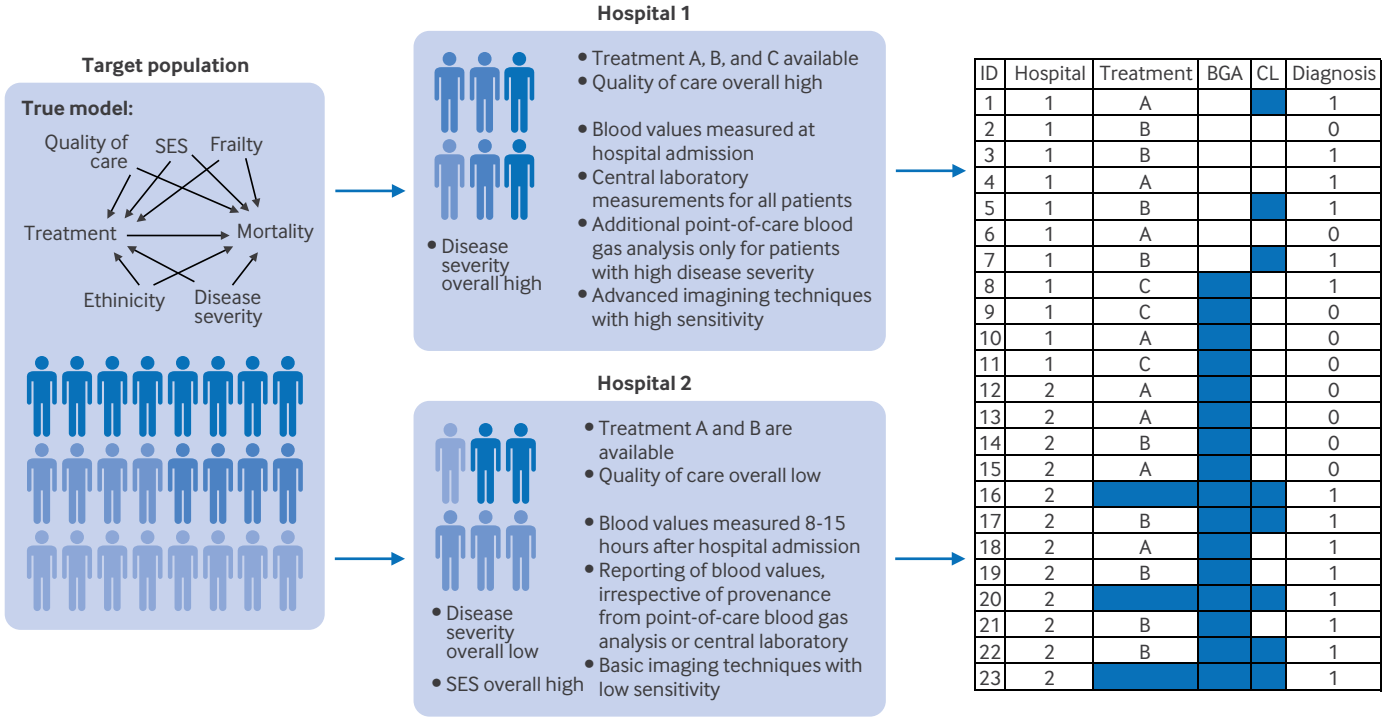
To evaluate the representativeness of the sample, population data can be used to compare distributions of demographic and clinical variables. For multisite studies, reporting how patients' characteristics and outcomes vary across sites is important. Findings derived from a single centre's database may need

verification against broader samples to ensure that the results hold across different populations and healthcare settings, and different matching and weighting approaches are available to improve representativeness.[74 75]

For prediction models, evaluating model performance in the target application setting is very important.[76] Prediction models can never be considered fully "validated."[77] In this sense, using prediction models that are clinically useful requires infrastructure and personnel for maintenance, continuous monitoring of model performance, and retraining in case of distribution shift. With routinely collected data, this may often occur whenever documentation practices, diagnoses, treatment options, or guidelines change over time or across different sites. Importantly, when evaluating a prediction model, one must not only assess its overall predictive performance but also evaluate clinical utility,[78] including net benefit for different subgroups that might be misrepresented or underrepresented in routinely collected data.[79] Net benefit quantifies a model's clinical utility to support a specific treatment decision by subtracting the weighted frequency of false positives from that of true positives. The weight reflects the relative importance of false negatives and false positives to capture the consequences of using the model to support clinical decision making. Net benefit is typically compared with alternative strategies, including competing models and the default strategies of treating everyone or treating no one.

Data quality

As routinely collected data are not recorded to answer a specific research question, many relevant variables may be inadequately or inconsistently captured. Understanding the data collection process is thus of critical importance. Recording and documentation practices may vary between different clinical settings and sites and over time, perhaps as a function of incentives and of the overall workload of the personnel collecting the data.[80 81] This creates a high risk of non-ignorable missing data and spurious associations,[41] resulting from structural problems such as variation



ID	Hospital	Treatment	BGA	CL	Diagnosis
1	1	A			1
2	1	B			0
3	1	B			1
4	1	A			1
5	1	B			1
6	1	A			0
7	1	B			1
8	1	C			1
9	1	C			0
10	1	A			0
11	1	C			0
12	2	A			0
13	2	A			0
14	2	B			0
15	2	A			0
16	2				1
17	2	B			1
18	2	A			1
19	2	B			1
20	2				1
21	2	B			1
22	2	B			1
23	2				1

Fig 1 | Fictive example of a study using routinely collected data involving data collected in a tertiary referral hospital (hospital 1) and a primary care hospital (hospital 2). Quality of care, socioeconomic status (SES), frailty, race, and disease severity are all confounders that would need to be accounted for to obtain unbiased estimates of treatment effects when the aim is explanation. However, these variables are not directly measured or are unmeasured, leading to a high risk of residual confounding or confounding by indication. The situation is further complicated by the fact that treatment C is available only in the tertiary hospital where patients have overall higher disease severity (represented in dark blue). In the dataset combining the patients from the two hospitals, missing data (illustrated as blue cells) are highly predictive of mortality as point-of-care blood gas analysis (BGA) measurements are done only for patients with high disease severity in hospital 1 and central laboratory measurements are not recorded whenever the patient died before the values became available. When the aim is prediction, a flexible algorithm may find that the most important features to predict mortality are the brand of the machine that is used to do imagining (which differs across hospitals), having point-of-care BGA measurements (only done for patients with high disease severity in hospital 1), and having a missing value for central laboratory (CL) measurements (which may occur when a patient dies before the test results are available, making the recording of this value superfluous). The fact that the decision to do certain tests is not random but a decision based on clinical knowledge that is not recorded in the data, implicit problems in time point alignment, and apparent associations merely arising from differences in eligibility and data quality, can thereby lead to algorithms that show impressive predictive performance but that will have little clinical usefulness

Box 1: Evidence and published examples of challenges arising in the analysis of routinely collected data (RCD)

Representativeness

- Evidence shows that ethnic minorities, patients without medical coverage, and low income and rural populations are underrepresented, whereas women, older people, more educated patients, and patients with a greater burden of comorbidity are overrepresented in RCD.[22-29] If patients from underrepresented groups are present in the data, a substantial risk exists that they may be misrepresented as they receive fewer diagnostics tests and interventions,[26] and they are more likely to visit multiple institutions[26 30-33] and to be lost during data linkage,[34] leading to incomplete and fragmented patient records
- These problems are further compounded when datasets are combined. In a systematic review of public covid-19 radiographic datasets, for instance, Garcia Santa Cruz and colleagues found that many were at a high risk of inducing bias in prediction models, because they included radiographs from children under 5 years of age from Guangzhou, China, as controls[35]
- Identifying a representative sample from the target population may be further complicated by the fact that inclusion and exclusion criteria can often not be established perfectly. Vanasse and colleagues and Alistair and colleagues report strong variations in treatment prevalence of schizophrenia (ranging from 0.59% to 1.49%) and in-hospital mortality rates (ranging from 14.5% to 30.1%) when comparing different criteria to identify patients with schizophrenia and sepsis in RCD, respectively[36 37]

Data quality

- Data audits generally report substantial error rates in RCD
- In a data audit of an international HIV research network, Duda and colleagues reported error rates exceeding 10% in key study variables, including weight measurements, antiretroviral drugs, and laboratory data, at five of seven sites[38]
- In a manual chart review comparing the operative note with the operative log for laparoscopic appendectomies reported to the American College of Surgeons National Surgical Quality Improvement Program, Childers and Maggard-Gibbons report that for 10 of 11 cases for which the operative note had been coded as not using a retrieval bag, the operative log mentioned the use of a retrieval bag[39]
- Peng and colleagues found evidence for substantial undercoding in hospital discharge abstract database records,[40] with estimated sensitivity ranging from 18.6% and 44.9% for obesity and depression to 68.3% and 84.6% for hypertension and diabetes
- Bilimoria and colleagues and Schwarzkopf and colleagues report how site specific reporting can create spurious associations and bias for venous thromboembolism rates and sepsis, respectively[41 42]
- In a survey among almost 30 000 patients, Bell and colleagues found that 21.1% reported documentation errors in electronic health record ambulatory care notes with higher error rates being reported by older and sicker patients and 6.5% of errors reflecting notes written on the wrong patient[43]
- If a study involves data linkage, an inherent trade-off between false matches and missed matches exists,[44] introducing additional errors
- Even in cases in which the documentation of prescribed drugs is correct, treatment status may be unknown as several studies have found adherence rates in routine care to be much lower than in randomised controlled trials, with only around half of patients actually taking treatments as they are prescribed and adherence rates showing substantial variation as a function of age and other sociodemographic factors[45-50]

Time point alignment

- In a representative sample of reports of comparative non-randomised studies that assessed the effectiveness and/or safety of drug treatments, Yaacoub and colleagues found that most reports used RCD and that in 72% (143/200) of studies eligibility, treatment assignment, and start of follow-up were not aligned.[51] In an audit on data quality in an international HIV research network, Duda and colleagues found that treatment regimens and associated dates and the timings of laboratory measurements were especially prone to error,[38] with error rates of up to 56% (54/96) and 42% (26/62), respectively. In such a situation, an alignment of eligibility, treatment assignment, and follow-up is challenging—sometimes impossible—to achieve
- Lack of time point alignment is also common in prediction models. In a model to predict hypertension, for instance, the most important features included lisinopril, hydrochlorothiazide, enalapril maleate, amlodipine besylate, and losartan potassium, all of which are used to treat hypertension.[52] In line with this, Winter and colleagues found that in 90% (18/20) of cases of clinical deterioration events, clinical actions reflecting changes in patient status and/or first order reflecting escalation of care preceded alerts through the paediatric Rothman Index,[13] which is an automated warning system integrated in the electronic health record system

Interventions, tests, and other clinical actions are not random or pre-specified

- Bosco and colleagues found that no adjustment method resolved confounding by indication when they assessed breast cancer recurrence among women treated with adjuvant chemotherapy compared with women who did not receive adjuvant chemotherapy.[53] Similarly, observational studies reported increased mortality with digoxin treatment. Re-analyses of the randomized DIG trial showed the inability to control bias in observational treatment comparisons despite extensive statistical adjustments: (non-randomised) prescription of digoxin was an indicator of disease severity and worse prognosis, which could not be fully accounted for by covariate adjustments in the DIG trial in which patients were well characterised[54]
- Similar problems arise from interventions, diagnostic tests, and other clinical actions not being random when building prediction models. In the analysis of chest radiographs, AI algorithms have been reported to use information on whether a portable scanner had been used in the detection of pneumonia and on the presence of thoracic tubes in the detection of pneumothorax.[55 56] Winkler and colleagues found that surgical skin markings substantially altered predictions of an algorithm for melanoma detection from dermoscopic images.[57] Agniel and colleagues found that for many common laboratory tests, the moment at which the test was done was more predictive of survival than the test result itself.[7] In a recent investigation of algorithmic fairness, Yang and colleagues found that healthcare disparities arising from underdiagnoses may be exacerbated through prediction models, with consistently larger fairness disparities of a state-of-the-art foundational model compared with radiologists in chest radiograph diagnosis; the model showed systematic underdiagnosis in female, younger, and black patients[58]

(Continued)

Box 1: (Continued)

Multiplicity of possible analysis strategies

- The analytical variability in the analysis of RCD is illustrated in cases in which different teams of researchers studying the same research question on the same dataset use different sets of covariates and operationalisations of covariates, different outcome measures, and different exclusion and inclusion criteria and publish contradictory and unreplicable results,[59 60] sometimes even in the same year and the same journal.[61] Using dabigatran versus warfarin initiation in patients with atrial fibrillation as an example, Wang and colleagues found that published estimates of the hazard ratio for major bleeding with dabigatran compared with warfarin ranged from 0.56 to 1.58.[62] They found that a third of variation in results was related to the use of alternative study design and analysis parameters, in particular differences in outcome algorithms and criteria used to define follow-up

of documentation practices across sites, from measurements being taken only for certain patient populations (for example, with high disease severity), or from a combination of these two factors. Excluding patients with missing values can therefore lead to selection bias.

In an urgent clinical situation, for instance, documenting clinical measurements irrelevant to immediate patient care is rightly of low priority. A clinician may then leave non-mandatory fields blank, creating missing data. For mandatory fields, a clinician may fill in an arbitrary allowable value, leading to substantial measurement error. Similarly, data collection and reporting are typically non-standardised. When patients from different sites are included in the same study, some sites may use advanced imaging techniques, use highly sensitive test kits, and have high coding accuracy, whereas others may use more basic imagining techniques, use lower sensitivity test kits, and have low coding accuracy. Figure 1 shows how differences in representativeness, reporting patterns, and quality of care across sites can lead to distortions. Finally, data quality may change over time depending on adaptations of their purpose or staff changes.[82] This is particularly problematic when such changes are associated with a studied treatment or prognostic factors. The implementation of a quality assurance programme to improve surgical outcomes may, for instance, be accompanied by the requirement of collecting additional items and higher standardisation of data collection, potentially leading to bias when the effectiveness of the programme is evaluated.[83]

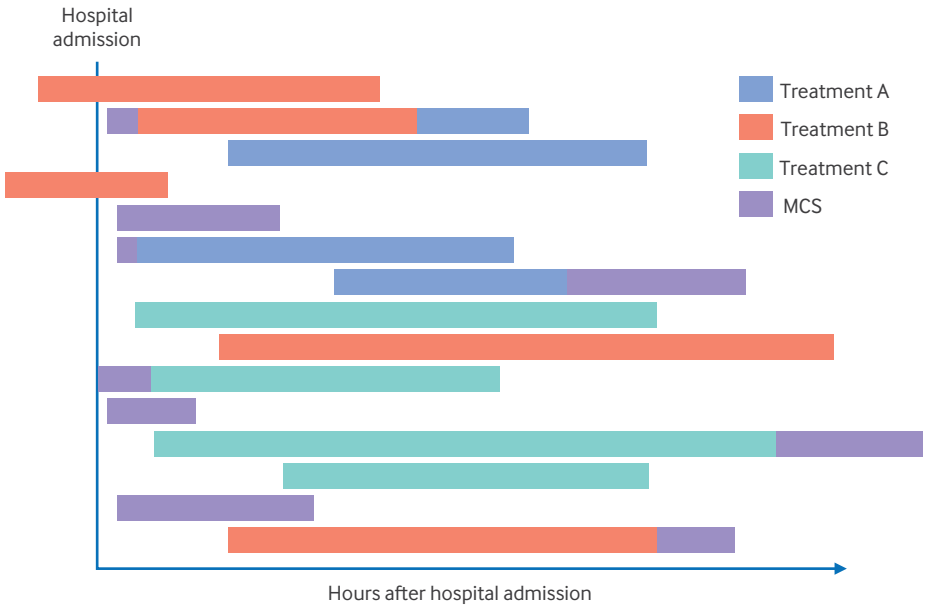
When the aim is explanation, measurement error and missing data in the exposure of interest and confounding variables may bias effect estimates, reduce statistical power, and distort functional forms.[84] Unsystematic measurement error in the outcome, on the other hand, will generally not lead to bias but only to a loss of statistical power.[84 85] In prediction, measurement error and informative missing data in predictor variables are of less concern if these aspects do not change when future predictions are made. In such a situation, a prediction model will simply incorporate the (possibly suboptimal) measured information relevant to predict future values, irrespective of how these measurements relate to what they purport to measure. However, if the aim is to generalise to other settings, measurement error and missingness arising from different measurement procedures will be more problematic.[86 87]

Tackling data quality

Studies using routinely collected data can be considered "high quality" only when sufficient steps to assess data quality, characterise potential data errors, identify reasons for missingness, and harmonise data across sites are undertaken.[88 89] In this sense, working with people who are familiar with how a certain dataset was collected and extracted is crucial, and the assessment of data quality is often context dependent. If the data are regularly used for research, standardised protocols for data collection and analysis enhance consistency and reduce variability across different sites. Training staff on data collection, ensuring adherence to predefined procedures, and code sharing minimise errors and improve overall quality of data. Validation studies, in which high quality data are prospectively collected simultaneously with routinely collected data, can provide information on patterns of missing data and on the characteristics and the magnitude of measurement error, which can then be adjusted for in statistical analyses.[90 91] As measurement error in the treatment of interest and in confounders is typically considered to be more problematic than that in the outcome,[84 85] conducting randomised controlled trials on routinely collected data,[92-94] in which treatments are randomly assigned and confounding is minimised through design, also represents a useful option to mitigate problems with data quality.

For explanation, the analyses should always be done using data that is of high quality. When the aim is prediction, on the other hand, improving data quality in the training data may be counterproductive when this leads to lower consistency of data quality between training and deployment. In this sense, extensive data cleaning is problematic when building prediction models if the same data cleaning will not be done once the prediction model is used.

Concerning missing data, it is typically important to consider mechanisms that may have led to missing values, assess and transparently report the proportion of missing data and predictors of missingness,[95 96] and account for missing data by using appropriate techniques. This last point will depend on whether the research aim is description, explanation, or



ID	A	B	C	MCS	Mortality
1	0	1	0	0	1
2	1	1	0	1	0
3	1	0	0	0	1
4	0	1	0	0	1
5	0	0	0	1	1
6	1	0	0	1	0
7	1	0	0	1	1
8	0	0	1	0	1
9	0	1	0	0	0
10	0	0	1	1	0
11	0	0	0	1	1
12	0	0	1	1	1
13	0	0	1	0	0
14	0	0	0	1	1
15	0	1	0	1	1

Fig 2 | Fictive example to illustrate problems that may commonly arise in time point alignment in routinely collected data. Hospital admission would be a natural time point at which patients are included, but treatment B is sometimes initiated before hospital admission whereas treatment initiation may be delayed for other patients. In this situation, an alignment of eligibility, treatment assignment, and start of follow-up is very difficult or even impossible to achieve. If the research question focuses on a comparison of the effectiveness of the three treatments in patients receiving mechanical circulatory support (MCS), this choice will lead to selection bias and severely bias treatment effects, because some patients receive it as bridge to recovery whereas it can be seen as a rescue therapy after the other treatments have failed in others. Most importantly, these problems are impossible to detect when the information on whether a treatment was received is reported as a binary variable (as seen in the right table) rather than as an event that has an associated start and stop date

prediction.[97] Doing sensitivity analyses is typically recommended to evaluate how findings depend on assumptions about measurement errors, misclassification, and missingness.

Time point alignment

When data are collected for research purposes, a natural alignment of eligibility, treatment assignment, and follow-up time points is usually present, because a natural "time zero," representing patient enrolment, often exists. Similarly, when building or evaluating prediction models on such data, a natural time point usually exists at which the prediction can be made, and which information is available at this time point is obvious. In routinely collected data, the first entry for a patient often has no clear clinical meaning (for example, beginning of the insurance) and previous diagnoses and treatments are often unknown. Patients often receive multiple interventions, and treatment status can best be described through a treatment sequence rather than a single treatment (fig 2). Data collection is influenced by daily clinical practice routines, leading to timings of different interventions and measurements being missing or misreported.[82] During a busy period, for instance, nurses may find time to record clinical events or changes in medication only at the end of their shift.

When the research aim is explanation, failure to align eligibility, treatment assignment, and start of follow-up gives rise to different biases, including problems that arise from the inclusion of prevalent users,[67 98] immortal time periods,[99] and selection bias due to post-treatment eligibility criteria.[100] The last occurs whenever patients are excluded if they received multiple treatments or when they are included in the treatment group only if they fill a specified number of consecutive prescriptions.[100 101]

Lack of time point alignment is a very common source of error when the aim is prediction.[102] Considered "one of the top ten data mining mistakes,"[103] time point misalignment leads to so-called data leakage, resulting in prediction models with impressive predictive performance and nearly no clinical value. In many situations, routinely collected data lack sufficient temporal resolution to make predictions actionable.[104] Even if temporal resolution is high, which time point to use is not always clear. For laboratory tests, for instance, the time point of test performance or the time point once results became available may be relevant. For explanation and description aims, the time point of test performance is relevant because it reflects the state of the patient at that time. For prediction, the time point at which results become available is relevant because it reflects knowledge on the state of the patient

and it is the time point when the measurement can first be used to make a prediction.

Tackling eligibility and time point alignment

Biases arising from inappropriate handling of different time points are sometimes regarded as self-inflicted,[67] as ensuring correct time point alignment is in principle possible. However, sensible alignment of time points in explanation and prediction is often challenging and requires substantial knowledge of the clinical background and data collection context. To improve transparency, an essential first step is to report and visualise the time points when inclusion and exclusion criteria are met, follow-up starts, and different key clinical variables are measured.[19 105] In prediction, the time when predictions are made should be indicated, along with the information available at this point. Marking all relevant variables on a timeline that is synchronised between patients by using a common index time (for example, disease onset or first hospital admission) makes problems with time point alignment more obvious. This allows improved communication between clinicians and analysts, who may not be aware of the exact timings of interventions and measurements.

If the aim is explanation, target trial emulation can be a helpful way to uncover and mitigate difficulties in time point alignment,[67] and several methods exist to correct for time point misalignment, including cloning, censoring and weighting,[106] and new user designs.[107] However, appreciating that target trial emulation does not overcome confounding by indication is important, and if the exact timing of different interventions is not recorded or is recorded incorrectly, aligning eligibility, treatment assignment, and start of follow-up may be impossible. As a consequence, from a conceptual point of view, the easiest solution to deal with this problem is arguably again through randomisation in a randomised controlled trial using routinely collected data.

Interventions and diagnostic tests are not random or pre-specified

When data are collected for research purposes, measurements and interventions are usually pre-specified and standardised to align with the research question. This typically involves obtaining measurements of all relevant outcomes (including adverse events) and of all other clinical variables that are relevant to the research question at regular and pre-specified time points. As routinely collected data are usually collected for a completely different purpose, interventions, diagnostic tests, and many other events are typically neither random nor pre-specified but are mainly steered by clinical decision making.

In the absence of randomised treatment assignment, ruling out residual confounding, or more precisely confounding by indication, in routinely collected data is nearly impossible. Confounding by indication arises whenever a treatment decision is influenced by factors that are unrecorded or imperfectly recorded, such as quality of care, resources at the study site, disease severity, socioeconomic status, and frailty. As a consequence, statistical techniques to overcome confounding often prove insufficient to eliminate this bias.[8 53 108] Even under the somewhat unrealistic assumption that adjustment for confounding is possible in observational studies using routinely collected data, estimating a clear treatment policy effect will often be very difficult, because no information on intended interventions exists, only information on actual interventions. In line with this, Hemkens and colleagues found that treatment effects estimated in observational studies using routinely collected data were substantially higher than those in subsequent randomised controlled trials for the same research questions.[109] At the same time, estimating the treatment effect for patients who actually take the treatment may also be impossible, because adherence rates in routine practice are in general lower than in randomised controlled trials and adherence is typically not closely monitored.[110] Differences in treatment adherence might to some extent explain why estimated treatment effects have been found to be 20% less favourable in trials using routinely collected data than with traditional trials.[111]

In prediction, confounding is not generally considered to be a problem,[112] because the aim is not to make causal statements but merely to make accurate predictions. However, being aware that clinical variables that may influence treatment decisions, health outcomes, and other clinical actions are often unmeasured is important. For instance, the probability of myocardial infarction is much higher for a patient with chest pain if a test for troponin has been ordered.[113] Similarly, patients who receive very frequent visits from family members in the intensive care unit may have a higher probability of dying in hospital.[30] Although these two variables may show a strong association with the respective outcomes, neither of them produces clinically useful predictions. Once physicians and patients start following the prediction model rather than the clinical knowledge that is not recorded in the data, these models will fail. In the two examples, a troponin test may not be ordered because the model predicts that the probability of myocardial infarction is low in the absence of a test order for troponin, and the family might not visit because the model predicts that the probability of the patient dying is low in the absence of frequent family visits.

This problem is relevant not only for predictors but also for the outcome of interest if it is not rigorously defined and measured. If the health outcome to be predicted is a diagnosis or clinical procedure, the outcome itself is in part the result of a clinical decision. Diagnoses as outcomes may be based on a test result that is known only when a certain test has been done. In this situation, one is faced with so-called partial verification bias,[114] whereby the disease status is unknown in patients with no test performed. Importantly, a large body of evidence shows that sociodemographic factors influence diagnoses

and clinical decisions.[115-117] In this situation, flexible prediction models will learn to reproduce differential actions that replicate and exacerbate these biases,[10 23 73] which include underdiagnosis bias for historically underserved populations.[118] Partial verification bias is also very relevant when the aim is explanation. For instance, known side effects of a treatment may be monitored more closely and ascertained with more detail for patients receiving this treatment. This differential reporting can thereby artificially inflate differences in the prevalence of these side effects for different treatments.

Tackling interventions and diagnostic tests not being random or pre-specified

When the aim is explanation, the most promising option to tackle confounding by indication is randomisation in randomised controlled trials using routinely collected data. However, conducting randomised controlled trials using routinely collected data remains challenging, as diagnostic tests are not pre-specified and, as a consequence, random treatment assignments do not resolve verification bias and outcome misclassification. Without randomisation, a range of techniques can be used to tackle confounding, ranging from standardisation, regression modelling, and various matching and weighting approaches to doubly robust estimation and machine learning methods.[119 120] Goetghebeur and colleagues provide a comprehensive introduction to formulating causal questions for observational studies and provide an overview of estimation approaches.[119] These approaches all have advantages and disadvantages and are suitable only for specific situations. Even for the most sophisticated statistical or machine learning techniques, the validity of results often depends on the assumption that all confounding variables have been measured perfectly, which is unlikely with routinely collected data. To assess sensitivity to unmeasured confounding, VanderWeele and Ding proposed E values,[121] which have recently gained in popularity in the analysis of observational studies.[122] However, they are criticised for not providing additional information and for making strong and often implausible assumptions that may result in misleading conclusions.[123 124] Quantitative bias analysis, instrumental variable approaches, falsification endpoints, and active comparator, new user designs can to some extent evaluate and adjust for bias arising from unmeasured confounding in observational studies using routinely collected data.[107 125-130] However, these approaches also have limitations with quantitative bias analysis requiring criteria to judge when the bias is too high. Also, in practice, suitable instruments for instrumental variable approaches are often lacking. Finally, falsification endpoints can show that an estimated effect is not valid. Contrarily, finding no effect for a falsification endpoint does not provide any proof that an estimated effect on routinely collected data is unbiased.

Multiplicity of possible analysis strategies

Recent work has increased awareness about multiple possible strategies in the analysis of empirical data.[131 132] Pressure to publish noteworthy results (that are clinically plausible) may result in selective reporting among the results arising from the multiplicity of possible analysis strategies, leading to overestimated treatment effects or overoptimistic measures of predictive performance that would not be replicated in subsequent studies.

Analytical variability in routinely collected data will be much greater than in the analysis of data that have been collected for research purposes, owing to a lack of knowledge of how the data were generated and control over how they were collected.[133] Researchers typically have great flexibility in the choice of their research hypotheses, and they have to choose among many imperfectly measured proxies of clinically relevant variables. Which patients to include in the analysis and how to treat missing values and outliers are often unclear. Even in the absence of incentives to produce significant results, researchers will have to select one among many possible analytical pathways.[62]

Tackling the multiplicity of possible analysis strategies

Prediction models that are chosen in a data driven way require suitable analysis strategies to avoid overfitting.[134] To obtain valid and reliable statistical inference, refraining from data driven variable selection is equally important when the aim is explanation.[135 136] In their review on comparative non-randomised studies assessing the effectiveness and/or safety of drug treatments, Yaacoub and colleagues found that most studies did not justify variables that had been included in the model (62%; 115/186) or used data driven variable selection (26%; 52/200).[51] A more suitable strategy for explanatory research using routinely collected data is to use clinical knowledge. This includes knowledge on mechanistic causal relations and evidence from the literature to construct and register a directed acyclic graph.[70] Ideally, this should be specified by independent experts masked to variables measured in the routinely collected data. A second step should evaluate whether all relevant confounders identified by the directed acyclic graph are available in the routinely collected data. Specifying all measured and unmeasured confounders and potential colliders and mediators is important to define an appropriate set of adjustment variables and to assess the extent of unmeasured and residual confounding, because some of the relevant confounders, including the frailty of the patient, socioeconomic status, and disease severity, will be measured imprecisely or not at all.

Naudet and colleagues argue that if studies on routinely collected data are to be considered as a source of information in regulatory or public policy decisions, the same basic quality standards as for interventional studies are needed.[133] Preregistering and publishing study design and analysis plans helps to avoid

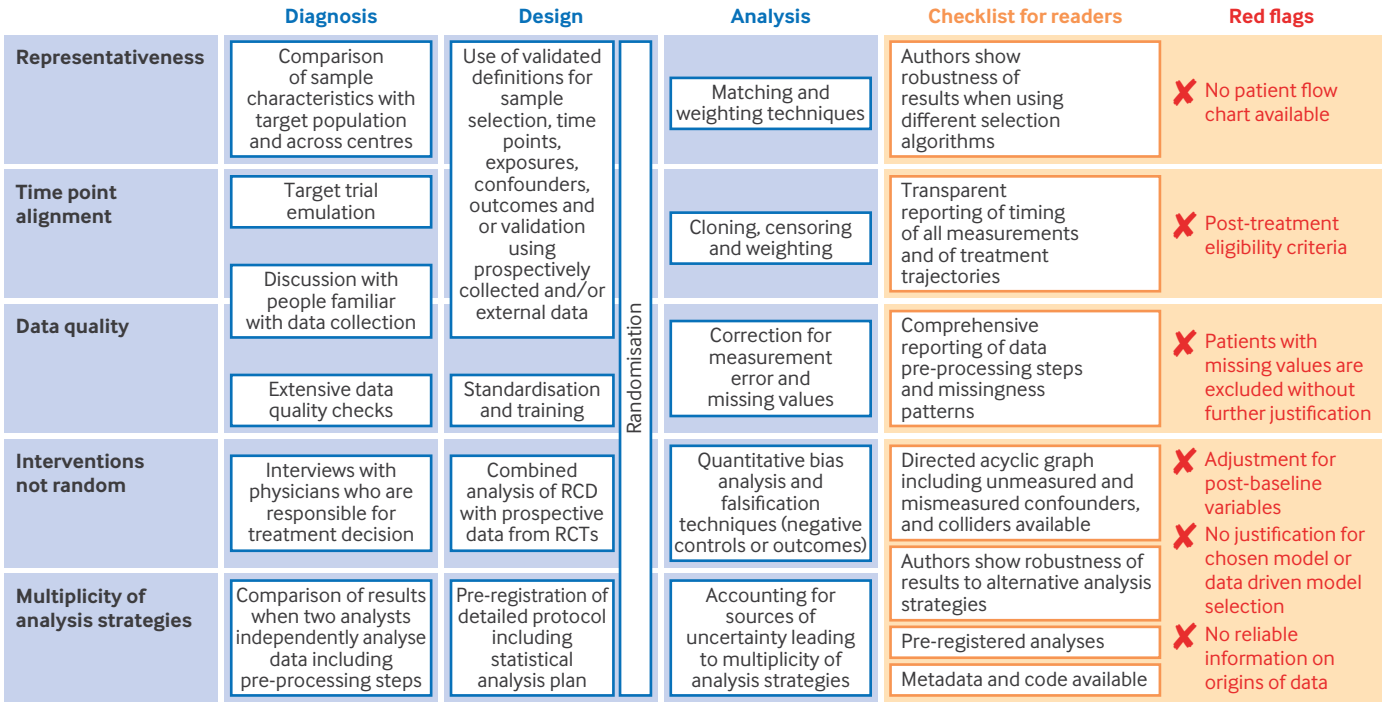


Fig 3 | Strategies to diagnose problems arising in the analysis of routinely collected data (RCD), with concrete steps across design and analysis that authors can take to improve the reliability of studies, alongside a checklist for critical appraisal by readers and a list of red flags. RCT=randomised controlled trial

selective reporting and ensures transparency.[137] Given the many analytical choices, registration for studies using routinely collected data is both more important and more challenging than for studies in which data are collected for research purposes. Which aspects of the dataset were known before specifying the analysis should be clear, and deviations, which are to be expected, should be transparently reported in a dedicated supplement. Authors may often find that registering all analysis steps is not feasible owing to large uncertainties in the analysis of routinely collected data. In this situation, helpful strategies may be to specifically register contingency plans to tackle problems in the analysis, to do a pilot study in which the feasibility and the methodology of the research project are assessed,[138] or to do a blind analysis in which analysis choices are made without information on the outcome.[139]

Good statistical practice and randomisation will often reduce analytical variability in the analysis of routinely collected data to some extent. Moreover, transparently reporting the results from multiple, often equally defensible, analysis choices in multiverse analysis or vibration of effects analyses can provide evidence on the robustness of results and help to identify the most important choices that lead to variability in results.[140-144] To understand and estimate the variability of results under more naturalistic conditions, so-called multi-analyst studies,[145 146] in which different teams of researchers answer the same research question by using the same dataset, are a promising option. Finally, advanced statistical models can be used to account for the sources of uncertainty that lead to the multiplicity of analysis strategies.[147]

Given the complexities of studies using routinely collected data, precise and transparent reporting is both more important and more challenging than for studies that prospectively collect data for research purposes. In light of the many options for data pre-processing, studies using routinely collected data should include a detailed report of all initial data analysis steps as an appendix to improve transparency and reproducibility.[88 148] Compared with other steps that can be taken to improve the quality of retrospective studies, reporting is a point that is relatively easy. Reporting should be tailored to the research aim and specific study setting,[149-151] and various reporting guidelines can be found on the website of the EQUATOR network (https://www.equator-network.org/), including for observational and interventional studies using routinely collected data (for example, RECORD-PE).[19 20 152]

Finally, in light of growing concerns associated with so-called paper mills and high impact cases of research fraud, including the surgisphere and the Eliezer Masliah scandals,[153-156] following FAIR principles with detailed meta-data seems crucial if individual participant data cannot be shared owing to privacy protection, as does sharing analysis code if this is possible. Figure 3 provides an overall roadmap with concrete steps for doing analyses that produce more reliable estimates with details on how to diagnose problems for each of the discussed challenges and how to counter them with concrete steps in the design

and analysis of routinely collected data when the aim is explanation. It also provides a checklist and a list of red flags for readers. Supplementary table B provides an overview of minimal requirements and acceptable and ideal approaches to tackle challenges that arise in the analysis of routinely collected data when the aim is explanation.

Can AI overcome the challenges in routinely collected data?

AI is a somewhat amorphous umbrella term that encompasses tasks such as learning, reasoning, problem solving, and perception, enabling machines to automatically derive insights from data, to make informed decisions, and to solve complex problems. A very important subfield of AI is machine learning, which focuses on constructing (often predictive) models on the basis of available data without the need for explicit programming. Machine learning has historically focused on constructing low error, potentially black box, models. Meanwhile, statistics was more concerned with valid inference and accurate representation of uncertainty. However, despite separate development of these fields, no conceptual difference exists between machine learning and statistics regarding the principles of fitting such models.

Used properly, AI algorithms have the potential to mitigate key challenges in routinely collected data. Digitisation may reduce health disparities and improve the representativeness of routinely collected data by reducing costs and providing broader access to care.[24] Data quality could be improved with AI assistance in data collection, extraction,[157] cleaning,[158] and validation.[21] When it comes to explanation, causal machine learning offers robustness to model misspecification, requiring fewer modelling decisions and flexibly capturing non-linear associations and interactions in the estimation of treatment effects without the need to pre-specify these associations.[159 160] As AI algorithms are powerful in detecting patterns in large amounts of heterogeneous data, they offer unprecedented opportunities when trained on large datasets of high quality.

However, many models are built on large, heterogeneous, multisite data, involving no previous filtering of predictor variables and no or only minimal pre-processing or communication between model developers and clinicians. Many of these features can make such algorithms highly vulnerable to biases that commonly arise in the analysis of routinely collected data. As these algorithms are designed to detect complex associations, the risk that they make "shortcuts,"[161] in which they use spurious features and idiosyncrasies in a routinely collected dataset that may often arise through problems in study design, rather than using clinically meaningful differences in pathology, is high. As illustrated in figure 1, associations may simply arise because specialised hospitals with higher disease prevalence do a different set of routine tests or may use more advanced imaging techniques. Moreover, without filtering predictors on the basis of clinical expertise, avoiding time point misalignment is difficult and the risk that clinician initiated actions will be included in prediction models is increased, leading to overoptimistic predictive performance and models that are "looking over clinicians' shoulders,"[113] rather than providing predictions with clinical utility.

Owing to their flexibility in the modelling of interactions and non-linear associations, flexible AI algorithms are more likely to pick up differential treatments and diagnoses on the basis of sociodemographic factors and to perpetuate disparities.[10 58 118] Although a logistic regression model may also be prone to such biases, it usually requires explicit formulation of a prediction model to include an interaction term between some measure of disease severity and race, sex, or sociodemographic status. Inclusion of such a term by researchers should give them pause to consider what this means from a clinical perspective. These problems may arise even when some measure of race, for instance, is not explicitly recorded, because, in contrast to clinicians, flexible algorithms have been shown to be capable of inferring race from other measures, such as medical imaging.[58 162] Similar problems may arise when a flexible algorithm is used to estimate individual treatment effects on claims data, for instance. Given that treatment adherence is often lower in younger patients and in those with low socioeconomic status,[45 46] the algorithm may incorrectly deem a treatment to be non-effective in these groups of patients, simply because the treatment is not taken as prescribed. Similarly, a treatment may seem to be safe in certain groups of patients because side effects are underdiagnosed. Findings on apparent treatment inefficiency or safety may then in turn wrongly discourage practitioners from prescribing or, on the contrary, wrongly encourage them to prescribe the treatment for these groups of patients in the future.

Finally, owing to the black box nature of many AI algorithms, many of the problems arising from short cuts, time point misalignment, clinician initiated data, and differential diagnoses may go largely unnoticed. Although the field of explainable and interpretable AI has made considerable progress, it is still less established and scientifically validated when it comes to making general prediction models easily interpretable in practice. This makes rigorous internal and external validation even more important for flexible AI algorithms.[163] Systematic investigation of these matters will prove complicated when models are trained across distributed datasets in federated learning and AI algorithms are proprietary.[10 164] Without rigorous internal and external validation, and in the absence of clinical and methodological expertise, AI as a tool may do more harm than good when analysing routinely collected data.

Conclusion

Ensuring quality and rigour in the analysis of routinely collected data is essential to generate trustworthy insights that improve patient care. Clinical knowledge,

methodological expertise, and information about data collection and generation, as well as critical thinking, are indispensable in the analysis and interpretation of routinely collected data. Consequently, generating high quality evidence from routinely collected data requires wide interdisciplinary collaboration, involving expertise in clinical research, epidemiology, computer science, statistics, and machine learning. Moreover, studies using routinely collected data should not be done carelessly or in a rush. Lack of data integrity may mean that prospectively collecting research data or randomly assigning treatments in a randomised controlled trial using routinely collected data if possible, or even not studying a particular question, would be better. Simply doing the best we can using routinely collected data is not good enough. No evidence for a research question may be preferable to evidence from a low quality routinely collected data study, as results may otherwise be misleading and overinterpreted. When challenges and limitations are carefully tackled, studies on routinely collected data can provide a valuable source of information and may complement, and sometimes pave the way for, studies with active data collection.

AUTHOR AFFILIATIONS

[1]Department of Statistics, Ludwig Maximilian University, Munich, Germany

[2]MRC Clinical Trials Unit at UCL, London, UK

[3]Institute for Medical Information Processing, Biometry and Epidemiology, Faculty of Medicine, LMU Munich, Munich, Germany

[4]Munich Center for Machine Learning (MCML), Munich, Germany

[5]Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria

[6]Department of Epidemiology, Care and Public Health Research Institute, Maastricht University, Netherlands

[7]Department of Development and Regeneration, KU Leuven, Leuven, Belgium

[8]Leuven Unit of Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium

[9]Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany

[10]Biostatistics Unit, Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA

[11]Department of Medical Statistics, University Medical Center Goettingen, Germany

[12]Institute for Quality and Efficacy in Health Care, Cologne, Germany

[13]University of Rennes CHU Rennes, Inserm, EHESP, IRSET (Institut de recherche en santé, environnement et travail), Rennes, France

[14]Institut Universitaire de France (IUF), Paris, France

[15]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, USA

[16]Institute of Nutrition and Food Sciences (IEL) -Nutritional Epidemiology, University of Bonn, Bonn, Germany

[17]Medizinische Klinik und Poliklinik II, Universitätsklinikum Bonn, Bonn, Germany

[18]Department of Cardiac Surgery, University Hospital Bonn, Bonn, Germany

[19]Department of Cardiology/Angiology, University Heart Center Frankfurt, Goethe University Hospital, Frankfurt, Germany

[20]Klinik für Innere Medizin I, Universitätsklinikum Jena, Jena, Germany

[21]Mid-German Heart Center, Department of Cardiology and Intensive Care Medicine, University Hospital, Halle (Saale), Germany

[22]Department of Cardiology, Cardiology I, University Medical Center Mainz, Mainz, Germany

[23]Department of Cardiology and Vascular Medicine, West German Heart and Vascular Center, University Hospital Essen, Essen, Germany

[24]Division of Cardiology, Pulmonology, and Vascular Medicine, Medical Faculty, University Hospital Düsseldorf, Düsseldorf, Germany

[25]Department III of Internal Medicine, Heart Center, University of Cologne, Cologne, Germany

[26]Department of Cardiology and Angiology, Hannover Medical School, Hannover, Germany

[27]Department of Cardiology, Heart Center Leipzig at University of Leipzig, Leipzig, Germany

Contributors: SH, HT, and EL contributed to the concept, supervision, and administrative, technical, or material support. SH, TM, MH, BB, and EL drafted the manuscript. All authors contributed to revision of the manuscript and approved the final draft. EL and HT contributed equally to the manuscript as last authors. EL is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: None.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Dissemination to participants and related patient and public communities: We will disseminate information from this paper to researchers, healthcare providers, and the public through presentations and traditional and social media.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Makady A, de Boer A, Hillege H, Klungel O, Goettsch W, on behalf of GetReal Work Package 1). What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. *Value Health* 2017;20:858-65. doi:10.1016/j.jval.2017.03.008

2 Pacheco RL, Martimbianco ALC, Riera R. Let's end "real-world evidence" terminology usage: A study should be identified by its design. *J Clin Epidemiol* 2022;142:249-51. doi:10.1016/j.jclinepi.2021.11.013

3 von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007;370:1453-7. doi:10.1016/S0140-6736(07)61602-X

4 Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *Chance* 2019;32:42-9doi:10.1080/09332480.2019.1579578.

5 Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. *BMJ* 2024;386:e078276. doi:10.1136/bmj-2023-078276

6 Bica I, Alaa AM, Lambert C, van der Schaar M. From Real-World Patient Data to Individualized Treatment Effects Using Machine Learning: Current and Future Methods to Address Underlying Challenges. *Clin Pharmacol Ther* 2021;109:87-100. doi:10.1002/cpt.1907

7 Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 2018;361:k1479. doi:10.1136/bmj.k1479

8 Collins R, Bowman L, Landray M, Peto R. The Magic of Randomization versus the Myth of Real-World Evidence. *N Engl J Med* 2020;382:674-8. doi:10.1056/NEJMsb1901642

9 Pörteners B, Jung C, Meyfroidt G. Trash in/trash out? Using routinely collected clinical data for data science in the ICU: Con. *Intensive Care Med* 2025;51:382-4. doi:10.1007/s00134-024-07739-3

10 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447-53. doi:10.1126/science.aax2342

11 Childers CP, Maggard-Gibbons M. Same Data, Opposite Results?: A Call to Improve Surgical Database Research. *JAMA Surg* 2021;156:219-20. doi:10.1001/jamasurg.2020.4991

12 Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern Med* 2021;181:1065-70. doi:10.1001/jamainternmed.2021.2626

13 Winter MC, Kubis S, Bonafide CP. Beyond Reporting Early Warning Score Sensitivity: The Temporal Relationship and Clinical Relevance of "True Positive" Alerts that Precede Critical Deterioration. *J Hosp Med* 2019;14:138-43. doi:10.12788/jhm.3066

14 Meng XL. Statistical Paradises and Paradoxes in Big Data (i): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election. *Ann Appl Stat* 2018;12:685-726doi:10.1214/18-AOAS1161SF.

15 Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science* 2014;343:1203-5. doi:10.1126/science.1248506
16 Bradley VC, Kuriwaki S, Isakov M, Sejdinovic D, Meng XL, Flaxman S. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* 2021;600:695-700. doi:10.1038/s41586-021-04198-4
17 The Lancet. Rethinking research and generative artificial intelligence. *Lancet* 2024;404:1.
18 Suchak T, Aliu AE, Harrison C, Zwiggelaar R, Geifman N, Spick M. Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database. *PLoS Biol* 2025;23:e3003152. doi:10.1371/journal.pbio.3003152
19 Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ* 2021;372:m4856. doi:10.1136/bmj.m4856
20 Kwakkenbos L, Imran M, McCall SJ, et al. CONSORT extension for the reporting of randomised controlled trials conducted using cohorts and routinely collected data (CONSORT-ROUTINE): checklist with explanation and elaboration. *BMJ* 2021;373:n857. doi:10.1136/bmj.n857
21 Kotecha D, Asselbergs FW, Achenbach S, et al, Innovative Medicines Initiative BigData@Heart Consortium, European Society of Cardiology, CODE-EHR international consensus group. CODE-EHR best practice framework for the use of structured electronic healthcare records in clinical research. *BMJ* 2022;378:e069048. doi:10.1136/bmj-2021-069048
22 Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. *Am J Epidemiol* 2016;184:847-55. doi:10.1093/aje/kww112
23 Mihan A, Pandey A, Van Spall HGC. Mitigating the risk of artificial intelligence bias in cardiovascular care. *Lancet Digit Health* 2024;6:e749-54. doi:10.1016/S2589-7500(24)00155-9
24 Mihan A, Van Spall HGC. Interventions to enhance digital health equity in cardiovascular care. *Nat Med* 2024;30:628-30. doi:10.1038/s41591-024-02815-z
25 Al-Sahab B, Leviton A, Loddenkemper T, Paneth N, Zhang B. Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview. *J Healthc Inform Res* 2023;8:121-39. doi:10.1007/s41666-023-00153-2
26 Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol* 2021;21:234. doi:10.1186/s12874-021-01416-5
27 Harton J, Mitra N, Hubbard RA. Informative presence bias in analyses of electronic health records-derived data: a cautionary note. *J Am Med Inform Assoc* 2022;29:1191-9. doi:10.1093/jamia/ocac050
28 Miller S, Wherry LR. Health and Access to Care during the First 2 Years of the ACA Medicaid Expansions. *N Engl J Med* 2017;376:947-56. doi:10.1056/NEJMsa1612890
29 Reddy H, Joshi S, Joshi A, Wagh V. A Critical Review of Global Digital Divide and the Role of Technology in Healthcare. *Cureus* 2022;14:e29739. doi:10.7759/cureus.29739
30 Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern Med* 2018;178:1544-7. doi:10.1001/jamainternmed.2018.3763
31 Arpey NC, Gaglioti AH, Rosenbaum ME. How Socioeconomic Status Affects Patient Perceptions of Health Care: A Qualitative Study. *J Prim Care Community Health* 2017;8:169-75. doi:10.1177/2150131917697439
32 Ng JH, Ye F, Ward LM, Haffer SC, Scholle SH. Data On Race, Ethnicity, And Language Largely Incomplete For Managed Care Plan Members. *Health Aff (Millwood)* 2017;36:548-52. doi:10.1377/hlthaff.2016.1044
33 Ramoni M, Sebastiani P. Robust Learning with Missing Data. *Mach Learn* 2001;45:147-70doi:10.1023/A:1010968702992.
34 Bohensky MA, Jolley D, Sundararajan V, et al. Data linkage: a powerful research tool with potential problems. *BMC Health Serv Res* 2010;10:346. doi:10.1186/1472-6963-10-346
35 Garcia Santa Cruz B, Bossa MN, Sölter J, Husch AD. Public Covid-19 X-ray datasets and their impact on model bias - A systematic review of a significant problem. *Med Image Anal* 2021;74:102225. doi:10.1016/j.media.2021.102225
36 Vanasse A, Courteau J, Fleury MJ, Grégoire JP, Lesage A, Moisan J. Treatment prevalence and incidence of schizophrenia in Quebec using a population health services perspective: different algorithms, different estimates. *Soc Psychiatry Psychiatr Epidemiol* 2012;47:533-43. doi:10.1007/s00127-011-0371-y
37 Johnson AEW, Aboab J, Raffa JD, et al. A Comparative Analysis of Sepsis Identification Methods in an Electronic Database. *Crit Care Med* 2018;46:494-9. doi:10.1097/CCM.0000000000002965
38 Duda SN, Shepherd BE, Gadd CS, Masys DR, McGowan CC. Measuring the quality of observational study data in an international HIV research network. *PLoS One* 2012;7:e33908. doi:10.1371/journal.pone.0033908
39 Childers CP, Maggard-Gibbons M. Re: Does retrieval bag use during laparoscopic appendectomy reduce postoperative infection?*Surgery* 2019;166:127-8. doi:10.1016/j.surg.2019.01.019
40 Peng M, Southern DA, Williamson T, Quan H. Under-coding of secondary conditions in coded hospital health data: Impact of co-existing conditions, death status and number of codes in a record. *Health Informatics J* 2017;23:260-7. doi:10.1177/1460458216647089
41 Bilimoria KY, Chung J, Ju MH, et al. Evaluation of surveillance bias and the validity of the venous thromboembolism quality measure. *JAMA* 2013;310:1482-9. doi:10.1001/jama.2013.280048
42 Schwarzkopf D, Rose N, Fleischmann-Struzek C, et al. Understanding the biases to sepsis surveillance and quality assurance caused by inaccurate coding in administrative health data. *Infection* 2024;52:413-27. doi:10.1007/s15010-023-02091-y
43 Bell SK, Delbanco T, Elmore JG, et al. Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes. *JAMA Netw Open* 2020;3:e205867. doi:10.1001/jamanetworkopen.2020.5867
44 Harron K. Data linkage in medical research. *BMJ Med* 2022;1:e000087. doi:10.1136/bmjmed-2021-000087
45 Fernandez-Lazaro CI, García-González JM, Adams DP, et al. Adherence to treatment and related factors among patients with chronic conditions in primary care: a cross-sectional study. *BMC Fam Pract* 2019;20:132. doi:10.1186/s12875-019-1019-3
46 Bristow RE, Chang J, Ziogas A, Anton-Culver H, Vieira VM. Spatial analysis of adherence to treatment guidelines for advanced-stage ovarian cancer and the impact of race and socioeconomic status. *Gynecol Oncol* 2014;134:60-7. doi:10.1016/j.ygyno.2014.03.561
47 Sabaté E. *Adherence to Long-term Therapies: Evidence for Action.* World Health Organization, 2003.
48 Ge L, Heng BH, Yap CW. Understanding reasons and determinants of medication non-adherence in community-dwelling adults: a cross-sectional study comparing young and older age groups. *BMC Health Serv Res* 2023;23:905. doi:10.1186/s12913-023-09904-8
49 Falagas ME, Zarkadoulia EA, Pliatsika PA, Panos G. Socioeconomic status (SES) as a determinant of adherence to treatment in HIV infected patients: a systematic review of the literature. *Retrovirology* 2008;5:13. doi:10.1186/1742-4690-5-13
50 Briesacher BA, Andrade SE, Fouayzi H, Chan KA. Comparison of drug adherence rates among patients with seven different medical conditions. *Pharmacotherapy* 2008;28:437-43. doi:10.1592/phco.28.4.437
51 Yaacoub S, Porcher R, Pellat A, et al. Characteristics of non-randomised studies of drug treatments: cross sectional study. *BMJ Med* 2024;3:e000932. doi:10.1136/bmjmed-2024-000932
52 Chiavegatto Filho A, Batista AFM, Dos Santos HG. Data Leakage in Health Outcomes Prediction With Machine Learning. Comment on "Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning". *J Med Internet Res* 2021;23:e10969. doi:10.2196/10969
53 Bosco JLF, Silliman RA, Thwin SS, et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *J Clin Epidemiol* 2010;63:64-74. doi:10.1016/j.jclinepi.2009.03.001
54 Aguirre Dávila L, Weber K, Bavendiek U, et al. Digoxin-mortality: randomized vs. observational comparison in the DIG trial. *Eur Heart J* 2019;40:3336-41. doi:10.1093/eurheartj/ehz395
55 Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med* 2018;15:e1002683. doi:10.1371/journal.pmed.1002683
56 Rueckel J, Trappmann L, Schachtner B, et al. Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs. *Invest Radiol* 2020;55:792-8. doi:10.1097/RLI.0000000000000707
57 Winkler JK, Fink C, Toberer F, et al. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatol* 2019;155:1135-41. doi:10.1001/jamadermatol.2019.1735
58 Yang Y, Liu Y, Liu X, et al. Demographic bias of expert-level vision-language foundation models in medical imaging. *Sci Adv* 2025;11:eadq0305. doi:10.1126/sciadv.adq0305
59 de Vries F, de Vries C, Cooper C, Leufkens B, van Staa TP. Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: the General Practice Research Database. *Int J Epidemiol* 2006;35:1301-8. doi:10.1093/ije/dyl147

60 Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. *Proc 2nd Mach Learn Healthc Conf* 2017;68:361-76.
61 Childers CP, Maggard-Gibbons M. Same Data, Opposite Results?: A Call to Improve Surgical Database Research. *JAMA Surg* 2021;156:219-20. doi:10.1001/jamasurg.2020.4991
62 Wang SV, Sreedhara SK, Bessette LG, Schneeweiss S. Understanding variation in the results of real-world evidence studies that seem to address the same question. *J Clin Epidemiol* 2022;151:161-70. doi:10.1016/j.jclinepi.2022.08.012
63 Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Routinely collected data and comparative effectiveness evidence: promises and limitations. *CMAJ* 2016;188:E158-64. doi:10.1503/cmaj.150653
64 Sauer CM, Dam TA, Celi LA, et al. Systematic Review and Comparison of Publicly Available ICU Data Sets-A Decision Guide for Clinicians and Data Scientists. *Crit Care Med* 2022;50:e581-8. doi:10.1097/CCM.0000000000005517
65 Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230. doi:10.1186/s12916-019-1466-7
66 Sauer CM, Chen LC, Hyland SL, Girbes A, Elbers P, Celi LA. Leveraging electronic health records for data science: common pitfalls and how to avoid them. *Lancet Digit Health* 2022;4:e893-8. doi:10.1016/S2589-7500(22)00154-6
67 Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol* 2016;183:758-64. doi:10.1093/aje/kwv254
68 Degtiar I, Rose S. A Review of Generalizability and Transportability. *Annu Rev Stat Appl* 2023;10:501-24doi:10.1146/annurev-statistics-042522-103837.
69 Griffith GJ, Morris TT, Tudball MJ, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun* 2020;11:5749. doi:10.1038/s41467-020-19478-2
70 Feeney T, Hartwig FP, Davies NM. How to use directed acyclic graphs: guide for clinical researchers. *BMJ* 2025;388:e078226. doi:10.1136/bmj-2023-078226
71 Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med* 2021;385:283-6. doi:10.1056/NEJMc2104626
72 Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health* 2023;2:e0000278. doi:10.1371/journal.pdig.0000278
73 Ghassemi M, Nsoesie EO. In medicine, how do we machine learn anything real?*Patterns (N Y)* 2022;3:100392. doi:10.1016/j.patter.2021.100392
74 Downes M, Gurrin LC, English DR, et al. Multilevel Regression and Poststratification: A Modeling Approach to Estimating Population Quantities From Highly Selected Survey Samples. *Am J Epidemiol* 2018;187:1780-90. doi:10.1093/aje/kwy070
75 Learoyd A, Nicholas J, Hart N, Douiri A. Application of information from external data to correct for collider bias in a Covid-19 hospitalised cohort. *BMC Med Res Methodol* 2024;24:149. doi:10.1186/s12874-023-02129-7
76 Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 2024;384:e074819. doi:10.1136/bmj-2023-074819
77 Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med* 2023;21:70. doi:10.1186/s12916-023-02779-w
78 Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ* 2024;384:e074820. doi:10.1136/bmj-2023-074820
79 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565-74. doi:10.1177/0272989X06295361
80 Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *J Med Internet Res* 2018;20:e185. doi:10.2196/jmir.9134
81 Ni K, Chu H, Zeng L, Li N, Zhao Y. Barriers and facilitators to data quality of electronic health records used for clinical research in China: a qualitative study. *BMJ Open* 2019;9:e029314. doi:10.1136/bmjopen-2019-029314
82 Callahan A, Shah NH, Chen JH. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. *Ann Intern Med* 2020;172(Suppl):S79-84. doi:10.7326/M19-0873
83 Chiolero A, Santschi V, Paccaud F. Public health surveillance with electronic medical records: at risk of surveillance bias and overdiagnosis. *Eur J Public Health* 2013;23:350-1. doi:10.1093/eurpub/ckt044
84 Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective.* 2nd ed. Chapman and Hall/CRC, 2006: 488doi:10.1201/9781420010138.
85 Keogh RH, Shaw PA, Gustafson P, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1-Basic theory and simple methods of adjustment. *Stat Med* 2020;39:2197-231. doi:10.1002/sim.8532
86 Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH, Collaborators. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J Clin Epidemiol* 2020;119:7-18. doi:10.1016/j.jclinepi.2019.11.001
87 Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J Clin Epidemiol* 2019;105:136-41. doi:10.1016/j.jclinepi.2018.09.001
88 Huebner M, Vach W, le Cessie S, Schmidt CO, Lusa L, Topic Group "Initial Data Analysis" of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies, http://www.stratos-initiative.org). Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. *BMC Med Res Methodol* 2020;20:61. doi:10.1186/s12874-020-00942-y
89 Schmidt CO, Struckmann S, Enzenbach C, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC Med Res Methodol* 2021;21:63. doi:10.1186/s12874-021-01252-7
90 Shepherd BE, Han K, Chen T, et al. Multiwave validation sampling for error-prone electronic health records. *Biometrics* 2023;79:2649-63. doi:10.1111/biom.13713
91 Ehrenstein V, Hellfritzsch M, Kahlert J, et al. Validation of algorithms in studies based on routinely collected health data: general principles. *Am J Epidemiol* 2024;193:1612-24. doi:10.1093/aje/kwae071
92 Mc Cord KA, Al-Shahi Salman R, Treweek S, et al. Routinely collected data for randomized trials: promises, barriers, and implications. *Trials* 2018;19:29. doi:10.1186/s13063-017-2394-5
93 Hemkens LG. How Routinely Collected Data for Randomized Trials Provide Long-term Randomized Real-World Evidence. *JAMA Netw Open* 2018;1:e186014. doi:10.1001/jamanetworkopen.2018.6014
94 Nyberg K, Hedman P. Swedish guidelines for registry-based randomized clinical trials. *Ups J Med Sci* 2019;124:33-6. doi:10.1080/03009734.2018.1550453
95 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 Inform* 2018;6:e11. doi:10.2196/medinform.8960
96 Carpenter JR, Bartlett JW, Morris TP, Wood AM, Quartagno M, Kenward MG. *Multiple Imputation and its Application.* John Wiley, 2023doi:10.1002/9781119756118.
97 Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol* 2020;125:183-7. doi:10.1016/j.jclinepi.2020.03.028
98 Danaei G, Tavakkoli M, Hernán MA. Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins. *Am J Epidemiol* 2012;175:250-62. doi:10.1093/aje/kwr301
99 Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol Drug Saf* 2007;16:241-9. doi:10.1002/pds.1357
100 Nguyen VT, Engleton M, Davison M, Ravaud P, Porcher R, Boutron I. Risk of bias in observational studies using routinely collected data of comparative effectiveness research: a meta-research study. *BMC Med* 2021;19:279. doi:10.1186/s12916-021-02151-w
101 Pocock SJ, Smeeth L. Insulin glargine and malignancy: an unwarranted alarm. *Lancet* 2009;374:511-3. doi:10.1016/S0140-6736(09)61307-6
102 Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y)* 2023;4:100804. doi:10.1016/j.patter.2023.100804
103 Nisbet R, Elder J, Miner GD. *Handbook of Statistical Analysis and Data Mining Applications.* Academic Press, 2009.
104 Tomašev N, Harris N, Baur S, et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nat Protoc* 2021;16:2765-87. doi:10.1038/s41596-021-00513-5
105 Gatto NM, Wang SV, Murk W, et al. Visualizations throughout pharmacoepidemiology study planning, implementation, and reporting. *Pharmacoepidemiol Drug Saf* 2022;31:1140-52. doi:10.1002/pds.5529
106 Hernán MA. How to estimate the effect of treatment duration on survival outcomes using observational data. *BMJ* 2018;360:k182. doi:10.1136/bmj.k182

107 Her QL, Rouette J, Young JC, Webster-Clark M, Tazare J. Core Concepts in Pharmacoepidemiology: New-User Designs. *Pharmacoepidemiol Drug Saf* 2024;33:e70048. doi:10.1002/pds.70048
108 Giordano SH, Kuo YF, Duan Z, Hortobagyi GN, Freeman J, Goodwin JS. Limits of observational data in determining outcomes from cancer therapy. *Cancer* 2008;112:2456-66. doi:10.1002/cncr.23452
109 Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ* 2016;352:i493. doi:10.1136/bmj.i493
110 Osterberg L, Blaschke T. Adherence to medication. *N Engl J Med* 2005;353:487-97. doi:10.1056/NEJMra050100
111 Mc Cord KA, Ewald H, Agarwal A, et al. Treatment effects in randomised trials using routinely collected data for outcome assessment versus traditional trials: meta-research study. *BMJ* 2021;372:n450. doi:10.1136/bmj.n450
112 Ramspek CL, Steyerberg EW, Riley RD, et al. Prediction or causality? A scoping review of their conflation within current observational research. *Eur J Epidemiol* 2021;36:889-98. doi:10.1007/s10654-021-00794-w
113 Beaulieu-Jones BK, Yuan W, Brat GA, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?*NPJ Digit Med* 2021;4:62. doi:10.1038/s41746-021-00426-3
114 O'Sullivan JW, Banerjee A, Heneghan C, Pluddemann A. Verification bias. *BMJ Evid Based Med* 2018;23:54-5. doi:10.1136/bmjebm-2018-110919
115 Best MJ, McFarland EG, Thakkar SC, Srikumaran U. Racial Disparities in the Use of Surgical Procedures in the US. *JAMA Surg* 2021;156:274-81. doi:10.1001/jamasurg.2020.6257
116 Li S, Fonarow GC, Mukamal KJ, et al. Sex and Race/Ethnicity-Related Disparities in Care and Outcomes After Hospitalization for Coronary Artery Disease Among Older Adults. *Circ Cardiovasc Qual Outcomes* 2016;9(Suppl 1):S36-44. doi:10.1161/CIRCOUTCOMES.115.002621
117 Simoni AH, Frydenlund J, Kragholm KH, Bøggild H, Jensen SE, Johnsen SP. Socioeconomic inequity in incidence, outcomes and care for acute coronary syndrome: A systematic review. *Int J Cardiol* 2022;356:19-29. doi:10.1016/j.ijcard.2022.03.053
118 Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 2021;27:2176-82. doi:10.1038/s41591-021-01595-0
119 Goetghebeur E, le Cessie S, De Stavola B, Moodie EEM, Waernbaum I, the topic group Causal Inference (TG7) of the STRATOS initiative. Formulating causal questions and principled statistical answers. *Stat Med* 2020;39:4922-48. doi:10.1002/sim.8741
120 Keele L, Grieve R. So Many Choices: A Guide to Selecting Among Methods to Adjust for Observed Confounders. *Stat Med* 2025;44:e10336. doi:10.1002/sim.10336
121 VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med* 2017;167:268-74. doi:10.7326/M16-2607
122 Guo J, Wang T, Cao H, et al. Application of methodological strategies to address unmeasured confounding in real-world vaccine safety and effectiveness study: a systematic review. *J Clin Epidemiol* 2025;181:111737. doi:10.1016/j.jclinepi.2025.111737
123 Ioannidis JPA, Tan YJ, Blum MR. Limitations and Misinterpretations of E-Values for Sensitivity Analyses of Observational Studies. *Ann Intern Med* 2019;170:108-11. doi:10.7326/M18-2159
124 Sjölander A, Greenland S. Are E-values too optimistic or too pessimistic? Both and neither!*Int J Epidemiol* 2022;51:355-63. doi:10.1093/ije/dyac018
125 Greenland S. Relaxation Penalties and Priors for Plausible Modeling of Nonidentified Bias Sources. *Stat Sci* 2009;24:195-210doi:10.1214/09-STS291.
126 Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;43:1969-85. doi:10.1093/ije/dyu149
127 Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Stat Med* 2014;33:2297-340. doi:10.1002/sim.6128
128 Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 2010;21:383-8. doi:10.1097/EDE.0b013e3181d61eeb
129 Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci U S A* 2018;115:2571-7. doi:10.1073/pnas.1708282114
130 Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Curr Epidemiol Rep* 2015;2:221-8. doi:10.1007/s40471-015-0053-5
131 Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 2011;22:1359-66. doi:10.1177/0956797611417632
132 Hoffmann S, Schönbrodt F, Elsas R, Wilson R, Strasser U, Boulesteix AL. The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. *R Soc Open Sci* 2021;8:201925. doi:10.1098/rsos.201925
133 Naudet F, Patel CJ, DeVito NJ, et al. Improving the transparency and reliability of observational studies through registration. *BMJ* 2024;384:e076123. doi:10.1136/bmj-2023-076123
134 Steyerberg EW. Overfitting and Optimism in Prediction Models. In: Steyerberg EW, ed. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer International Publishing, 2019: 95-112doi:10.1007/978-3-030-16399-0_5.
135 Heinze G, Dunkler D. Five myths about variable selection. *Transpl Int* 2017;30:6-10. doi:10.1111/tri.12895
136 Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J* 2018;60:431-49. doi:10.1002/bimj.201700067
137 Van Calster B. Provide public access to ethics-approved study protocols. *Nature* 2023;614:227. doi:10.1038/d41586-023-00329-1
138 Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. *J Educ Eval Health Prof* 2013;10:12. doi:10.3352/jeehp.2013.10.12
139 MacCoun R, Perlmutter S. Blind analysis: Hide results to seek the truth. *Nature* 2015;526:187-9. doi:10.1038/526187a
140 Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency Through a Multiverse Analysis. *Perspect Psychol Sci* 2016;11:702-12. doi:10.1177/1745691616658637
141 Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology* 2008;19:640-8. doi:10.1097/EDE.0b013e31818131e7
142 Klau S, Schönbrodt F, Patel C, Ioannidis J, Boulesteix AL, Hoffmann S. Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology. *Meta-Psychol* 2023;7:MP.2020.2556.
143 Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol* 2015;68:1046-58. doi:10.1016/j.jclinepi.2015.05.029
144 Vinatier C, Hoffmann S, Patel C, et al. What is the vibration of effects?*BMJ Evid Based Med* 2025;30:61-5. doi:10.1136/bmjebm-2023-112747
145 Aczel B, Szaszi B, Nilsonne G, et al. Consensus-based guidance for conducting and reporting multi-analyst studies. *Elife* 2021;10:72185. doi:10.7554/eLife.72185
146 Silberzahn R, Uhlmann EL, Martin DP, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Adv Methods Pract Psychol Sci* 2018;1:337-56doi:10.1177/2515245917747646.
147 Rehms R, Ellenbach N, Rehfuess E, Burns J, Mansmann U, Hoffmann S. A Bayesian hierarchical approach to account for evidence and uncertainty in the modeling of infectious diseases: An application to COVID-19. *Biom J* 2024;66:e2200341. doi:10.1002/bimj.202200341
148 Heinze G, Baillie M, Lusa L, et al, TG2 and TG3 of the STRATOS initiative. Regression without regrets -initial data analysis is a prerequisite for multivariable regression. *BMC Med Res Methodol* 2024;24:178. doi:10.1186/s12874-024-02294-3
149 Benchimol EI, Smeeth L, Guttmann A, et al, RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015;12:e1001885. doi:10.1371/journal.pmed.1001885
150 Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378
151 Langan SM, Schmidt SAJ, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ* 2018;363:k3532. doi:10.1136/bmj.k3532
152 Cashin AG, Hansford HJ, Hernán MA, et al. Transparent Reporting of Observational Studies Emulating a Target Trial-The TARGET Statement. *JAMA* 2025;334:1084-93. doi:10.1001/jama.2025.13350
153 Sanderson K. Science's fake-paper problem: high-profile effort will tackle paper mills. *Nature* 2024;626:17-8. doi:10.1038/d41586-024-00159-9
154 Abalkina A, Aquarius R, Bik E, et al. 'Stamp out paper mills' - science sleuths on how to fight fake research. *Nature* 2025;637:1047-50. doi:10.1038/d41586-025-00212-1
155 Servick K, Enserink M. The pandemic's first major research scandal erupts. *Science* 2020;368:1041-2. doi:10.1126/science.368.6495.1041

156 Piller C. Picture imperfect. *Science* 2024;385:1406-12. doi:10.1126/science.adt3535
157 Hossain E, Rana R, Higgins N, et al. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. *Comput Biol Med* 2023;155:106649. doi:10.1016/j.compbiomed.2023.106649
158 Myhre PL, Tromp J, Ouwerkerk W, et al. Digital tools in heart failure: addressing unmet needs. *Lancet Digit Health* 2024;6:e755-66. doi:10.1016/S2589-7500(24)00158-4
159 Feuerriegel S, Frauen D, Melnychuk V, et al. Causal machine learning for predicting treatment outcomes. *Nat Med* 2024;30:958-68. doi:10.1038/s41591-024-02902-1
160 Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide-prediction, machine learning and causal inference. *Int J Epidemiol* 2021;49:2058-64. doi:10.1093/ije/dyz132
161 Banerjee I, Bhattacharjee K, Burns JL, et al. "Shortcuts" Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation. *J Am Coll Radiol* 2023;20:842-51. doi:10.1016/j.jacr.2023.06.025
162 Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health* 2022;4:e406-14. doi:10.1016/S2589-7500(22)00063-2
163 Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021;3:e745-50. doi:10.1016/S2589-7500(21)00208-9
164 Lijović L, Elbers P. Leveraging the power of routinely collected ICU data. *Intensive Care Med* 2025;51:163-6. doi:10.1007/s00134-024-07745-5

Web appendix: Supplementary tables