

Addressing confounders in observational comparative effectiveness research: Methods, software, and reporting standards

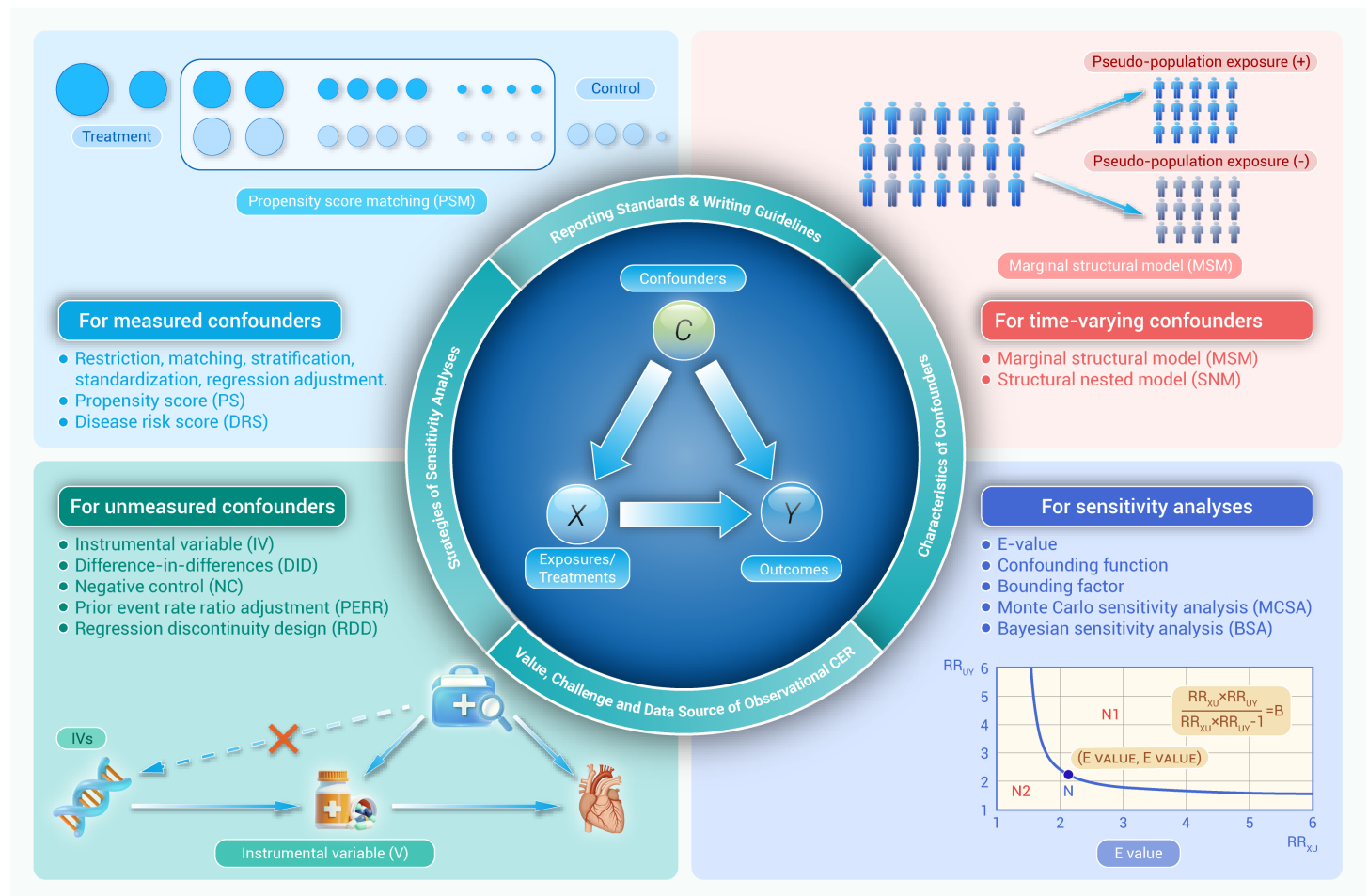
Yingying Wang,^{1,11} Wenwen Li,^{1,11} Ling Wang,² Yongyue Wei,^{3,4,5} Yang Zhao,^{6,7,8,9} Guoshuang Feng,^{10,*} and Lihong Huang^{1,*}

*Correspondence: huang.lihong@zs-hospital.sh.cn (L.H.); glxfgsh@163.com (G.F.)

Received: August 22, 2025; Accepted: December 31, 2025; Published Online: January 3, 2026; <https://doi.org/10.59717/j.xinn-med.2026.100187>

© 2026 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Confounding is the primary threat to validity in observational comparative effectiveness research (CER).
- Multiple statistical methods have been adopted to address measured, unmeasured, and time-varying confounders.
- Sensitivity analyses are essential to evaluate the robustness of findings in the presence of confounders.
- Writing guidelines and reporting standards are fundamental to ensure transparency and reproducibility.

Addressing confounders in observational comparative effectiveness research: Methods, software, and reporting standards

Yingying Wang,^{1,11} Wenwen Li,^{1,11} Ling Wang,² Yongyue Wei,^{3,4,5} Yang Zhao,^{6,7,8,9} Guoshuang Feng,^{10,*} and Lihong Huang^{1,*}

¹Department of Biostatistics, Zhongshan Hospital, Fudan University, Shanghai 200032, China

²Department of Health Statistics, Air Force Medical University, Xi'an 710032, China

³Center for Public Health and Epidemic Preparedness & Response, Peking University, Beijing 100191, China

⁴Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China

⁵Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China

⁶Department of Biostatistics, Nanjing Medical University, Nanjing 211166, China

⁷Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, China

⁸China International Cooperation Center for Environment and Human Health, Nanjing Medical University, Nanjing 211166, China

⁹Ministry of Education Key Laboratory for Modern Toxicology, School of Public Health, Nanjing Medical University, Nanjing 211166, China

¹⁰Big Data Center, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China

¹¹These authors contributed equally

*Correspondence: huang.lihong@zs-hospital.sh.cn (L.H.); glxfgsh@163.com (G.F.)

Received: August 22, 2025; Accepted: December 31, 2025; Published Online: January 3, 2026; <https://doi.org/10.59717/j.xinn-med.2026.100187>

© 2026 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Citation: Wang Y., Li W., Wang L., et al. (2026). Addressing confounders in observational comparative effectiveness research: Methods, software, and reporting standards. *The Innovation Medicine* 4:100187.

Confounding poses a critical threat to the validity of observational comparative effectiveness research (CER) by distorting the estimated associations between treatments/exposures and outcomes. This challenge is particularly pronounced in real-world data, given the absence of randomization and the prevalence of unmeasured and time-varying confounders. Thus, the observed associations may be attributable to differences other than the treatments/exposures of interest and causality cannot be assumed. By synthesizing theoretical foundations, practical applications, software implementations, and reporting standards, this study provides insights into statistical approaches for identifying and handling confounders in observational CER, in order to strengthen the validity of causal inference in real-world evidence, and ultimately enhance the quality and reliability of observational CER.

INTRODUCTION

Definition and impact of confounders

Confounding is typically illustrated as a common cause pathway between the treatments/exposures and the outcomes. Confounders refer to third variables that simultaneously influence both the treatments/exposures and the outcomes in a study. These factors (*C*) are characterized by being associated with the treatments/exposures (*X*), having an independent effect on the outcomes (*Y*) regardless of the treatment/exposure status, and not being mediators (*M*) in the causal pathway from the treatments/exposures to the outcomes. These key characteristics of confounders (*C*) can be intuitively presented using a directed acyclic graph (DAG), as shown in Figure 1. The presence of confounders can distort the estimated associations between the treatments/exposures and the outcomes, leading to potential overestimation or underestimation of the true effects.¹

Confounders in observational CER

Randomized controlled trials (RCTs) have widely served as the gold standard for comparative studies of the treatment/exposure effects. Their superiority lies in the ability to balance participant characteristics between groups and mitigate the influence of confounders through the process of randomization, thereby providing more reliable evidence of the causal relationship.² Unfortunately, RCTs have limitations such as enrolling homogeneous participants with low trial risks and high treatment responsiveness through strict inclusion and exclusion criteria, leading to a gap between trial settings and real-world clinical practice. Additionally, the high costs, long follow-up periods, ethical issues and recruitment challenges often render RCTs unfeasible to implement. Thus, observational CER can serve as valuable alternatives.

Observational CER was first proposed by the U.S. Agency for Healthcare Research and Quality (AHRQ) in 2009.³ They refer to the empirical analyses of

real-world treatment/exposure choices and outcomes to compare the practical effectiveness of different treatment/exposure regimens, often using data from electronic health records (EHRs), health insurance claims databases, disease registries, patient-reported outcome (PRO) surveys, and public health surveillance systems. As non-randomized controlled designs, observational CER can provide long-term follow-up data, be conducted in settings that more closely resemble real-world clinical practice, report on rare outcomes, and generate evidence for interventions that cannot be randomized. These attributes enable an opportunity to study the effectiveness of certain treatments/exposures in the real-world and confirm the results found in RCTs.⁴

However, the central threat to the validity of observational CER is confounders, which occur throughout the entire research process.⁵ During the period of study design, treatments/exposures in observational CER are not externally imposed by researchers but are determined by routine clinical practice, a process that introduces confounding by indication.⁶ For example, clinicians may prefer to prescribe the new anti-hypertensive drug to patients with more severe conditions, resulting in significantly higher baseline blood pressure levels and complication rates in the new drug group, thus interfering with efficacy evaluation. This practice leads to an imbalanced distribution of confounders between groups.⁶ During the period of implementation, treatment/exposure regimens are often adjusted dynamically according to symptoms or adverse reactions of patients. For example, a patient using oral anti-diabetic drugs may switch medications due to hypoglycemia, which can result in the failure to reflect the true efficacy of the original drugs. During the period of data analysis and result interpretation, data collection typically aligns with routine clinical practice rather than a predefined research protocol. Without proper adjustment, both measured and unmeasured confounders can distort the estimated treatment/exposure effects, undermining the validity of causal inferences drawn from observational CER.⁷ For example, failing to control cardiovascular disease history may overestimate the strength of the association between using Saxagliptin and incident acute myocardial infarction (AMI). This is because cardiovascular disease history not only influences treatment/exposure decisions, but also is inherently a risk factor for the outcome.⁸

In this study, we aim to provide a comprehensive overview of several methods to deal with confounders in observational CER. By synthesizing theoretical frameworks and practical applications, we seek to offer insights into statistical approaches for confounding handling in observational CER, in order to strengthen the validity of causal inference in real-world evidence, and ultimately enhance the quality and reliability of observational CER.

Statistical methods for dealing with confounders in observational CER

A wide series of statistical methods have been proposed and widely applied in observational studies.⁹⁻¹³ These methods can be broadly catego-

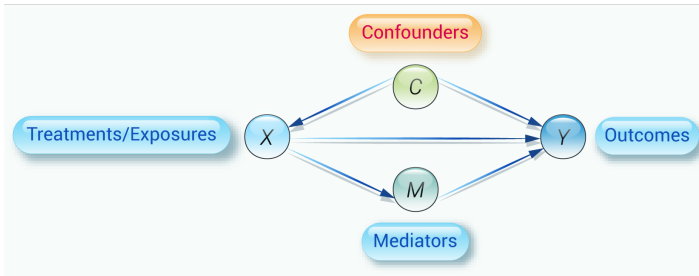


Figure 1. Causal diagram (DAG) of confounders In causal diagrams, confounders (C) distort the true effects of the treatments/exposures (X) on the outcomes (Y) by creating a non-causal backdoor pathway ($X \leftarrow C \rightarrow Y$). This makes X appear to affect Y even in the absence of a causal effect, or distorts the magnitude and direction of true effect estimates. In contrast, mediators (M) lie on the causal pathway ($X \rightarrow M \rightarrow Y$) and mediate the true effects of the treatments/exposures (X) on the outcomes (Y). While the M don't distort the total effect estimates, adjusting for them mistakenly blocks the actual causal effects of X on Y .

ized into four categories: methods for dealing with measured confounders, methods for dealing with unmeasured confounders, methods for conducting

sensitivity analyses, and methods for dealing with time-varying confounders. In the following sections, we outline key methods in each category, detailing their core principles, model specifications and illustrative case studies. Additional methods not discussed in detail are summarized in Table 1 for completeness.

STATISTICAL METHODS FOR DEALING WITH MEASURED CONFOUNDERS

Measured confounders refer to variables that can be directly obtained through research design planning or data collection processes, such as age, gender, history of underlying diseases, and laboratory test indicators. The core of their handling is to balance the distribution of such variables among different treatment/exposure groups using reasonable statistical methods to eliminate their interference with the treatment/exposure-outcome association, and sensitivity analyses must be conducted after handling to verify the robustness of the results.

Traditional methods

When addressing measured confounders, traditional statistical methods include restriction, matching, stratification, standardization, and regression

Table 1. Summary of statistical methods for dealing with confounders in CER.

Category	Methods	Conditions	Strengths	Limitations
Measured confounders	Restriction	Homogeneous subgroups; small sample size acceptable.	Simple implementation; eliminates confounding within strata.	Reduces sample size/generalizability; residual confounding possible.
	Matching	Sufficient overlap between groups; sufficient sample size for pairing.	Creates comparable groups by matching on confounders; balances observed covariates.	Fails on unmeasured confounders; difficult to match on multiple confounders; discards unmatched subjects.
	Stratification	Adequate sample size per stratum; confounder categories not too numerous.	Simple to interpret; allows assessment of treatment/exposure effects within different confounder strata.	Becomes complex with multiple confounders; sparse strata reduce power; residual within-stratum confounding possible.
	Standardization	Need for cross-population comparisons; reference population is defined.	Enables comparisons across different populations by adjusting to a common standard.	Choice of reference population affects results; may not capture all confounders.
	Regression Adjustment	Correct model specification (e.g., linear, logistic, Cox).	Efficiently handles multiple confounders simultaneously; uses full sample.	Prone to overfitting/collinearity; fails if unmeasured confounders exist.
	Propensity score matching (PSM)	Treatment/exposure assignment is influenced by confounders; overlap in PS distribution exists between groups.	Condenses multiple confounders into a single score; reduces dimensionality; balances observed covariates; mimics randomization.	Requires sufficient overlap in PS; discards unmatched individuals, reducing sample size; sensitive to PS model misspecification; matching quality affects results.
	Inverse probability treatment weighting (IPTW)	Treatment/exposure probabilities can be estimated; extreme weights can be truncated.	Retains all data, enhancing statistical power; explicitly balances confounders via weighting.	Susceptible to extreme weights (requires truncation); depends on accurate treatment/exposure probability estimation.
	PS stratification or subclassification	PS can be divided into meaningful strata (e.g., quintiles).	Retains all data; assesses treatment/exposure effects within strata and combines results.	Strata may not be homogeneous; pooling stratum effects can be complex; residual confounding possible.
	Covariate adjustment using the PS	PS can be included as a covariate in the outcome model; correct model specification (e.g., linear, logistic, Cox).	Simple implementation; can handle interaction terms.	Does not balance covariates across groups (unlike PSM/IPTW); may have biased variance estimation.
	Disease risk score (DRS)	Outcome risk can be estimated in the absence of treatment/exposure; historical or untreated data are available.	Condenses multiple confounders into a single score; useful for rare treatments/exposures or time-varying exposures.	Relies on accurate outcome risk estimation; performance declines with rare outcomes.

Table 1. (Continued)

Category	Methods	Conditions	Strengths	Limitations
Unmeasured confounders	Instrumental variable (IV)	Valid IV (associated with treatments/exposures, not directly with outcomes, independent of confounders of the treatment/exposure–outcome relationship); treatments/exposures are at least partially influenced by IV.	Addresses unmeasured confounding under strict IV assumptions; mimics randomization.	Difficult to find valid IVs; weak IVs cause bias; sensitive to IV assumptions.
	Difference-in-differences (DID)	Parallel trends assumption; no time-varying confounders.	Handles for time-fixed and group-specific confounders; suitable for policy or intervention evaluations.	Relies on parallel trends (difficult to verify); sensitive to time-varying confounders post-intervention.
	Negative control (NC)	Negative control exposure (NCE) or outcome (NCO) is available; shares confounding structure with primary analysis.	Detects residual confounding empirically; flexible (multiple NC types).	Limited by availability of valid negative controls.
	Prior event rate ratio adjustment (PERR)	Similar pre-treatment/exposure trends; no time-varying confounders.	Mitigates unmeasured confounders using pre-treatment/exposure data.	Assumes equal confounding pre/post; not always plausible.
	Regression discontinuity design (RDD)	Treatment/exposure assignment is determined by a cutoff on a continuous variable; individuals near the cutoff are comparable.	Exploits natural cutoff to mimic randomization near the threshold; strong internal validity.	Applies only to settings with clear cutoff points; external validity may be limited.
Sensitive analyses for unmeasured confounders	E-value	Observed treatment/exposure–outcome association is available (RR, OR, HR); no assumptions on distribution of confounders.	Quantifies the minimum unmeasured confounding strength needed to explain away the observed association; simple to compute.	Assumes symmetric confounding; does not estimate bias magnitude; relies on accurate effect estimates.
	Confounding function	Based on the counterfactual principle; confounders require no assumptions.	Quantifies bias from hypothetical confounders.	With binary intervention; relies on professional knowledge; complexity; strong subjectivity.
	Bounding factor	Confounders require no assumptions.	Bias ranges can be assessed via mathematical boundary derivation.	Insufficient precision; subjective boundary setting.
	Monte Carlo sensitivity analysis (MCSA)	Prior knowledge on distributions of confounders is available.	Flexibility in complex modeling; incorporates uncertainty.	Computationally intensive; relies on plausible parameter distributions.
	Bayesian sensitivity analysis (BSA)	Prior knowledge on distributions of confounders is available.	Integrates prior knowledge; full probabilistic framework.	Sensitivity to priors; complex implementation.
Time-varying confounders	Marginal structural model (MSM)	Time-varying exposures and confounders exist; confounders are influenced by prior exposures and affect future outcomes.	Handles dynamic confounding via time-varying inverse probability weights; mimics randomization over time.	Requires longitudinal data; sensitive to model misspecification in weight calculation; extreme weights may arise.
	Structural nested model (SNM)	Complex dynamic relationships between exposures, confounders, and outcomes; parametric form of the causal model can be specified.	Flexibly models dynamic interactions; incorporates prior knowledge of causal pathways.	Requires advanced statistical expertise; data-intensive; sensitive to model assumptions.

adjustment.^{14,15} Restriction refers to confining the study population to a subgroup where the confounders remain constant, like including only non-smokers in a study on lung disease. However, it may lead to a reduction in sample size and has limited extrapolation capability. Matching entails pairing treated/exposed and controlled participants based on confounders like age and sex to enhance group comparability, yet it may exclude unmatched individuals and reduce statistical efficiency.¹⁶ Stratification divides the study

population into distinct strata, such as different tumor stage groups, ensuring that each stratum is relatively homogeneous with respect to the specific characteristics, which is a straightforward method but becomes challenging with multiple confounders.¹⁷ Standardization creates a common standard population and then applies the observed rates or outcomes from each group to this standard population, making it valuable for cross-population comparisons, but the choice of reference population can influence the results.

Regression adjustment incorporates confounders as covariates in regression models. While offering flexibility, this method relies on correct model specification and is prone to overfitting and collinearity issues when handling multiple covariates.

However, in high-dimensional data modeling, the interactions between treatment/exposure variables and other covariates become extremely complex, and conventional methods may not be sufficient to handle them. To mitigate this issue, some approaches like the propensity score (PS) and the disease risk score (DRS) have been widely used in observational CER to reduce the model complexity and effectively handle multiple measured confounders.

Propensity score (PS)

Propensity score (PS) was first proposed by Rosenbaum and Rubin in 1983.¹⁸ It is defined as the probability that an individual is assigned to a certain treatment/exposure ($X=1$), given a set of observed baseline covariates (Z). Essentially, the core idea of PS is to condense the combined influence of multiple covariates into a single comprehensive score.¹⁹ Adjusting for the PS can effectively balance the distribution of these covariates between the treatment/exposure group and control group in observational CER, commonly through four methods including matching (PS matching, PSM), weighting (inverse probability of treatment weighting, IPTW), stratification (PS stratification/subclassification), and covariate adjustment. This process achieves an effect akin to randomization, also known as "post hoc randomization".

Model specification. The PS is typically estimated using logistic regression models, with the covariates (Z) as independent variables and the certain treatment/exposure variables (X , where $X = 1$ denotes the treatment/exposure group and $X = 0$ denotes the control group) as dependent variables. This model estimates the probability for each participant of being assigned to the treatment/exposure group ($X=1$). The resulting probability value is PS, which ranges between 0 and 1.

PS matching (PSM). Propensity score matching (PSM) constructs comparable treatment/exposure group and control group by selectively pairing participants with similar PS value, aimed at achieving "quasi-randomization" baseline balance through individual pairing.²⁰ A critical prerequisite for PSM is the overlap assumption that both groups share a common range of PS values. Two common matching algorithms are nearest neighbor matching and caliper matching. Nearest neighbor matching operates by selecting controls with the closest PS values for each treated/exposed individual. Caliper matching restricts pairing to those where the absolute difference in PS values falls within a predefined caliper width. This caliper width is mostly set to 0.2 standard deviations (SD) of the overall PS distribution, as this setting balances matching accuracy and sample retention rate. If the sample size is sufficient, the caliper width can be narrowed (e.g., 0.1 SD) to further enhance group balance. If the sample size is scarce, the caliper width needs to be relaxed or 1:k ($k \leq 4$) matching strategy adopted. The matching process inherently excludes unmatched participants, redefining the study population to those with overlapping PS distributions. While this enhances comparability by enforcing covariate balance, it reduces statistical power and estimate precision compared to weighting or covariate adjustment methods, primarily due to loss of sample size.²¹

Inverse probability treatment weighting (IPTW). Inverse probability treatment weighting (IPTW) uses PS to balance covariates between treatment/exposure group and control group, by assigning each participant a weight that is the inverse probability of receiving the actual treatment/exposure.²² Formally, weights are calculated as $1/PS$ for the treated/exposed participants and $1/(1-PS)$ for the controls. Conceptually, IPTW can be considered mathematically equivalent to standardization. Compared with matching method, weighting preserves the entire sample size to enhance statistical power and enable explicit specification of the target of inference through selection of different weights.²³ However, critical implementation challenges involve extreme weights, resulting in some participants having a very high or low probability of being treated/exposed. If extreme weights are truncated, the representativeness of the sample may be compromised.²⁴

PS stratification/subclassification. PS stratification/subclassification typically subclassifies the participants into several strata based on quintiles

of PS.²⁵ Within each stratum, the distribution of the covariates is expected to be similar between groups. The overall treatment/exposure effects are derived by combining the stratum-specific effects through weighted averaging techniques, with two predominant schemes involving weighting by the proportion of subjects in each stratum for population-level treatment/exposure effects, and weighting by inverse variance to optimize precision. Despite retaining all samples for the analysis, PS stratification/subclassification method presents critical limitations, such as the complexity of pooling the strata effects, the decrease of effective sample size after stratification and the existence of unbalanced residual confounding.²⁶

Covariate adjustment using the PS. The PS can be adjusted as a single covariate into the regression model for estimating the treatment/exposure effects.²⁶ The primary advantage of this method lies in its capacity to comprehensively integrate multiple covariates and their interaction terms within the PS calculation, making the model more parsimonious. However, the PS-based covariate adjustment differs from other PS methods. Rather than balancing covariates across groups, this methodology primarily deals with confounding through statistical adjustment in the outcome model. Additionally, the conventional variance estimation of the treatment/exposure effects may be biased.²⁷

Practical application scenarios and usage trends of four PS methods. In observational CER, the selection of PS methods is determined by study objectives, data structure, and clinical practice requirements. Varga et al. summarized the usage trends of these four PS methods in 84 observational studies on measured confounders.⁹ PSM is the most widely used, accounting for 56% (47/84) of studies. This method is particularly adapted for short-term outcome comparison studies of two regimens with a moderate sample size, such as efficacy comparisons between different surgical therapies and etiological exploration of rare adverse events (AEs). IPTW follows as 36% (30/84), and is especially applicable to long-term outcome studies using large-scale real-world data such as EHRs and health insurance claims databases, or scenarios where the sample size differs significantly between groups. PS stratification/subclassification is used in 31% (26/84) of studies, and is mainly suitable for clinical studies that need to identify populations with superior benefits, such as risk stratification analysis in oncology or chronic disease researches, or scenarios with a large sample size. Covariate adjustment using PS is the least independently applied, at 15% (13/84). This method is mainly suitable for small-sample exploratory studies or multiple method verification scenarios.

Case study. Sanket et al. explored the associations of risks for in-hospital mortality and major bleeding with the use of intravascular microaxial left ventricular assist device (LVAD) versus intra-aortic balloon pump (IABP) among patients with acute myocardial infarction (AMI) complicated by cardiogenic shock, using data from two registries.²⁸ Given that LVAD and IABP are not randomly assigned in clinical practice, their use often depends on the baseline characteristics of patients. Thus, this study adopted the method of PSM to handle measured confounders. Specifically, a total of 1680 patients who received LVAD were 1:1 matched with 1680 patients who received IABP based on factors like demographics and medical history. After matching, the standardized mean difference (SMD) for 74 out of 75 baseline characteristics was less than 0.1, showing good balance between the groups.

Disease risk score (DRS)

Disease risk score (DRS) was first proposed by Miettinen in 1976.²⁹ It is defined as the probability of an outcome occurring ($Y=1$) for an individual who has not received a certain treatment/exposure or been exposed to a certain factor ($X=0$), given a set of observed baseline covariates (Z). The estimation of DRS is independent of the treatment/exposure variables, making it applicable to scenarios involving rare exposures or those with multiple exposure categories and levels. Generally, the impact of confounders on disease risk is more stable over time compared to their impact on the propensity to be treated/exposed. This stability implies that if confounders are adequately measured and balanced by the DRS at baseline, any subsequent change in disease risks within the treatment/exposure group can be attributed to the treatment/exposure variables themselves.

Model specification. DRS is commonly estimated using logistic regression models, with data source from the full cohort (FC), untreated/unexposed

participants only (UO),³⁰ historical data (HD) or alternate data (AD).³¹ The estimation procedure comprises two stages: modeling fitting and DRS estimation. Notably, in the DRS method, the data used for model fitting and that for DRS estimation are typically derived from different datasets, or even distinct samples.

1. Based on FC: All observations from the FC are included in model fitting, with the covariates (Z) and the treatment/exposure variables (X) as independent variables, as well as the outcome variables (Y , where $Y = 1$ denotes the occurrence of outcome and $Y = 0$ denotes non-occurrence) as dependent variables to estimate the probability for each participant of occurring outcome (Y). In DRS estimation stage, for any individual in the FC, the treatment/exposure variables (X) are set to zero (where $X = 0$ denotes the untreated/unexposed condition), then the FC data is substituted back into the previous model to obtain the DRS estimate, which ranged between 0 and 1.

2. Based on UO, HD or AD: Samples without treatment/exposure from UO, HD or AD data are included in modeling fitting, with the covariates (Z) as independent variables and the outcome variables (Y , where $Y = 1$ denotes the occurrence of outcome and $Y = 0$ denotes non-occurrence) as dependent variables to estimate the probability of occurring an outcome (Y) for each participant. In DRS estimation stage, the FC data is substituted back into the previous model, to obtain the DRS estimate, which ranged between 0 and 1.

DRS vs. PS. Like PS, DRS can be applied in multiple analytical frameworks, including matching, stratification, or as a continuous covariate alongside treatment/exposure variables in statistical models. Although both DRS and PS aim to summarize measured confounders into a single comprehensive score, their underlying methodological logics are distinct. The PS estimates the probability of each participant being assigned to a certain treatment/exposure based on covariates, while the DRS estimates the probability of each participant developing an outcome of interest based on covariates.³² Both PS and DRS have limitations. PS is limited when treatment/exposure is rare or changes over time. By contrast, DRS is limited when outcome is rare.³⁰ For example, for newly marketed drugs, both the number of users and the occurrence of drug-related outcomes are relatively rare. In the early post-marketing phase, the propensity for individuals to use the new drug may change significantly over time and can be influenced by other covariates, leading to instability. These factors limit the applicability of the traditional PS and DRS methods in post-marketing studies evaluating the safety and effectiveness of new drugs. Thus, the methods of high-dimensional propensity score (HDPS) and high-dimensional disease risk score (HD-DRS) can be used to overcome this problem.^{33,34} It should be noted that the implementation of HDPS and HD-DRS methods inherently relies on the integration of machine learning algorithms.

Case study. Sengwee et al. evaluated the risk of acute myocardial infarction (AMI) associated with Saxagliptin in new users, leveraging real-world EHR data from the United States.⁸ Given the complexity of confounders in real-world data, including factors that influence usage of Saxagliptin and heterogeneity in baseline risks for AMI, both DRS stratification and PSM were applied to adjust for measured confounders. Specifically, the Saxagliptin users were matched with comparators with similar PS values to balance baselines. The DRS for new users in the surveillance cohort was calculated by multiplying their covariate profiles by regression coefficients derived from an earlier sample. Participants were stratified by DRS to align baseline AMI risk between groups. This dual approach minimized multi-source confounding, enhancing the validity of conclusions on Saxagliptin and AMI risk.

STATISTICAL METHODS FOR DEALING WITH UNMEASURED CONFOUNDERS

Unmeasured confounders refer to variables that cannot be obtained due to data collection limitations (e.g., socioeconomic status and details of living habits not recorded in electronic health records) or research design flaws (e.g., failing to consider genetic susceptibility). Their handling is more difficult than measured confounders, and the core is to quantify their potential impact through specific methods or evaluate the tolerance of results to unmeasured confounders through sensitivity analyses.

Instrumental variable (IV)

To address unmeasured confounders, instrumental variable (IV) method

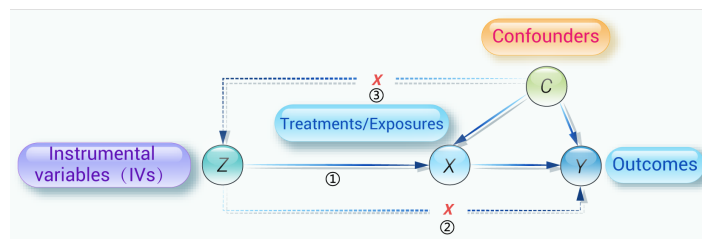


Figure 2. Causal diagram (DAG) of instrumental variables (IVs) The IVs (Z) affect the treatments/exposures (X) (condition ①) and affects the outcome (Y) only indirectly via the treatments/exposures (X) (condition ②). Simultaneously, the independence of IVs (Z) from the confounders (C) blocks the confounding pathway from C to Y (condition ③). This allows for the isolation and estimation of the pure causal effects of X on Y by leveraging the exogenous variation in X induced by Z , thus effectively addressing the bias introduced by unmeasured confounding.

has been widely adopted. It was first proposed by PG. Wright in 1928 and initially applied in the fields of economics and sociology.³⁵ This approach addresses the endogeneity issue by introducing IVs that are correlated with the endogenous variable but uncorrelated with the error term, thereby yielding consistent parameter estimates. IV method requires the use of valid IVs that satisfy three critical conditions: (1) the IVs (Z) are statistically associated with the treatments/exposures (X), (2) the IVs (Z) influence the outcomes (Y) exclusively through the treatments/exposures (X), with no direct effect, and (3) the IVs (Z) are independent of all confounders (C) of the treatment/exposure-outcome relationship (Figure 2).³⁶

Model specification. The two-stage least squares (2SLS) method is used to estimate the causal effects of the endogenous treatment/exposure variables (X) on the outcome variables (Y) by leveraging instrumental variables (Z). The procedure consists of the following two stages. In the first stage, a linear regression model is fitted with the treatment/exposure variables (X) as the dependent variables and the instrumental variables (Z) as independent variables, to obtain the estimators of treatment/exposure variables ($\hat{X}$). For the instrumental variables (Z) to be valid, they must satisfy the conditions of being statistically associated with the treatment/exposure variables (X) ($\text{Cov}(Z, X) \neq 0$) and being independent of the error term ($\text{Cov}(Z, e) = 0$) in the model. In the second stage, the outcome variables (Y) are regressed on the estimator of treatment/exposure variables ($\hat{X}$) obtained from the first stage, aiming to isolate the causal effects of X on Y by accounting for the role of Z .³⁷

IV vs. MR. The IV method functions similarly to random assignment. Specifically, random assignment is independent of all confounders, but is associated with the actual treatment/exposure assignments, and affects the outcomes only through the treatments/exposures. In fact, the random assignment variables are IVs. In recent years, Mendelian Randomization (MR) has been widely applied in etiological analyses and disease prognosis studies.^{38,39} This method utilizes genetic variants, such as single nucleotide polymorphisms (SNPs), as IVs for causal inference.⁴⁰ The rationale lies in the fact that genetic variants are randomly allocated during gamete formation and conception in accordance with Mendel's laws. This random allocation enables genetic variants to serve as instruments for randomly dividing the study population into groups with different levels of the exposures under study. When the sample size is sufficiently large and the genetic variants are not associated with potential confounders, the quasi-random allocation should result in groups that are, on average, comparable in all aspects except for levels of the exposure.⁴¹ Such characteristics underscore that the core value of the IV method lies in providing causal evidence comparable to that from RCTs. However, its most prominent limitation is to identify the valid and qualified IVs.

Recent advance. To relax the stringent assumptions required by traditional IV method, recent advances in causal inference under latent confounders have introduced new concepts such as the conditional instrumental variable (CIV) and ancestral instrumental variable (AIV).^{42,43} A CIV relaxes the assumptions of a standard IV by conditioning on a set of measured variables. However, a variable may be a CIV when it has zero causal effect on exposure, and this might result in a misleading conclusion. To address this limitation, AIV in DAG is proposed by imposing more restrictive condition.⁴³ Thus, AIV, defined within a DAG, can be viewed as a

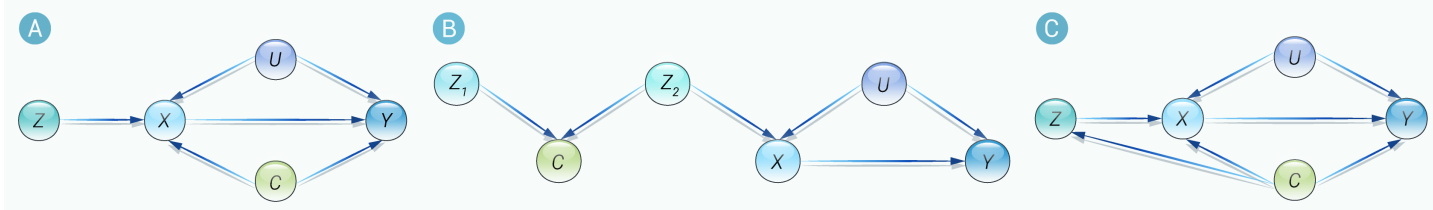


Figure 3. Causal diagram (DAG) of IV, CIV and AIV (a) Z is an IV for X (exposure) → Y (outcome). (b) Z₁ is a CIV with C (confounder) as its conditional set, but not an AIV; Z₂ is an AIV. (c) Z is both a CIV and an AIV conditional on C (confounder).

constrained form of a CIV that facilitates efficient identification of valid conditioning sets. In summary, an AIV in DAG is a CIV in DAG, but a CIV may not be an AIV. AIV is a restricted version of CIV (Figure 3). Building on this idea, some algorithms have been developed to identify an appropriate conditioning set Z for a given AIV within a maximal ancestral graph (MAG).⁴⁴ Furthermore, Cheng et al. developed a theory based on MAG to facilitate the search for candidate AIVs and to determine its corresponding conditioning set for each identified AIV.⁴⁵ While the IV framework is a powerful tool for addressing latent confounding, traditional IV techniques struggle to handle latent time-varying confounders in longitudinal studies. To address this challenge, Cheng et al. proposed the time-dependent instrumental factor model (TIFM), a novel framework for estimating time-varying causal effects in the presence of latent time-varying confounding.⁴⁶ At each time step, TIFM employs a recurrent neural network (RNN) architecture to infer latent IV factors. The inferred latent IVs are then incorporated into a 2SLS regression to estimate the average causal effect at each time point. For detailed descriptions of these methods, please refer to the original study.

Case study. David et al. investigated the association between the intraoperative fraction of inspired oxygen (FIO₂) and the risk of postoperative organ injury using a retrospective observational cohort design.⁴⁷ However, intraoperative FIO₂ levels are not randomly assigned but are adjusted by anesthesiologists based on patient-specific factors, such as underlying diseases and intraoperative vital signs, which may simultaneously influence both FIO₂ levels and the risk of postoperative organ injury, introducing unmeasured confounders. To address this, the typical FIO₂ levels each anesthesiologist tends to use in routine practice was set as an IV, satisfying three validity conditions. The preferences of anesthesiologists directly affect intraoperative FIO₂ levels, but do not directly cause postoperative organ injury. And random patient assignment to anesthesiologists makes these preferences independent of confounders like underlying diseases or surgical risk.

Difference-in-differences (DID)

Difference-in-differences (DID) method was first proposed by Ashenfelter in 1978 and has been widely used in the fields of economics for policy effect evaluation. In recent years, it has also been increasingly applied in observational studies across diverse domains including public health interventions, environmental regulations and social program assessments.⁴⁸⁻⁵⁰ This approach estimates the treatment/exposure effects by comparing changes in outcomes between the treatment/exposure group and the control group before and after a certain treatment/exposure, while handling for time-invariant confounders and baseline group differences. Specifically, there are two types of changes in the outcome variables: the changes in the treatment/exposure group ($\Delta_A = A_1 - A_0$) and the changes in the control group ($\Delta_B = B_1 - B_0$). The DID estimator is then computed as the difference between these changes: $\Delta_A - \Delta_B$, effectively removing time-varying effects common to both groups (Figure 4).⁵¹

Model specification. When applying the traditional DID model, three assumptions are critical for valid causal inference: (1) parallel trends assumption, the cornerstone of DID, positing that in the absence of treatment/exposure, the trends in outcomes would evolve identically for both treatment/exposure group and control group; (2) stable unit treatment value assumption (SUTVA) or no spillover effects assumption, holding that the potential outcomes of any unit depend only on its own treatment/exposure status; (3) no anticipation effect assumption, stipulating that units do not alter behavior prior to treatment/exposure implementation.^{52,53} The traditional regression model of DID includes a treatment/exposure term (X), a time term

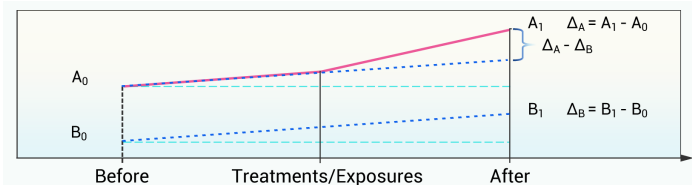


Figure 4. Diagram of difference-in-differences (DID) The solid line illustrates the post-treatment/exposure trends ($\Delta_A = A_1 - A_0$), and the dashed line illustrates the parallel pre-treatment/exposure trends ($\Delta_B = B_1 - B_0$). Here the DID estimator ($\Delta_A - \Delta_B$) captures the causal effect as the vertical difference between the actual treatment/exposure trajectory (Δ_A) and its counterfactual trend (Δ_B).

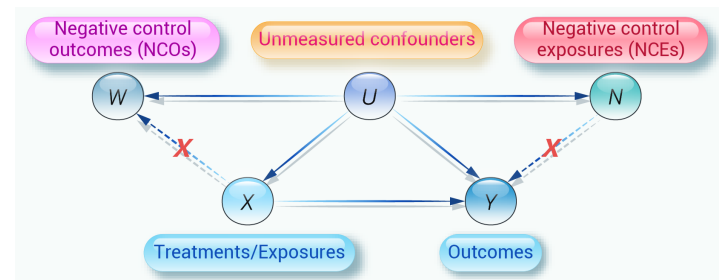


Figure 5. Causal diagram (DAG) of NCEs and NCOs The arrows indicate causal or associative pathways. The dashed lines from the treatments/exposures (X) to the NCOs (W), and from the NCEs (N) to the outcomes (Y) present that any observed association along these pathways is non-causal and arises solely due to unmeasured confounders (U). This diagram visually demonstrates how NCEs and NCOs serve as detective tools to identify and adjust for unmeasured confounding, thereby enhancing the validity of causal inference.

(T), and an interaction term between the treatments/exposures and the time ($X \times T$). The treatment/exposure term (X, where $X = 1$ denotes the treatment/exposure group and $X = 0$ denotes the control group) captures the inherent differences in the outcomes between the treatment/exposure group and control group that are constant over time. The time term (T) captures the aggregate time trends that similarly affect both groups from the pre- to post-treatment/exposure periods, accounting for external factors that change over time and affect both groups. The interaction term is the core of the DID estimator. The coefficient on this interaction term corresponds to the causal effects of the treatments/exposures.⁴⁹

Case study. Andrew et al. explored the associations of real-time continuous glucose monitoring (rtCGM) with glycemic control and acute metabolic events among patients receiving insulin therapy for diabetes.⁵⁴ The DID method was used to isolate the causal effects of rtCGM use by addressing inherent baseline differences and concurrent temporal trends between rtCGM users and non-users. The net effect of rtCGM exposure was then quantified as the changes in the rtCGM users from pre-treatment to post-treatment minus those in the non-users. This design enabled causal inference under the parallel trends assumption, wherein both groups shared identical outcome trajectories prior to rtCGM adoption.

Negative control (NC)

Control is a fundamental principle in clinical research design. Inspired by the rule that "negative controls would yield negative results in the absence of confounding", the negative control (NC) methods in epidemiological studies

have evolved as an extension of this idea. The NC is defined as a comparator that does not participate in the causal hypothesis of the study, yet shares the same potential confounding structure as the research hypothesis.⁵⁵

In 2010, Lipsitch et al. proposed the concepts of negative control exposure (NCE) and negative control outcome (NCO) to detect confounding and identify noncausal associations.⁵⁵ This approach has been elegantly applied to the discussion about the effects of vaccination of the elderly on pneumonia or influenza hospitalization and on all-cause mortality. Specifically, the NCEs (M) are defined as treatments/exposures that have no causal effect on the outcomes of interest (Y), but are associated with either unmeasured confounders (U) or the target treatments/exposures (X). Conversely, the NCOs (W) are defined as outcomes that experience no causal effect from the target treatments/exposures (X), but are linked to unmeasured confounders (U) (Figure 5). If any association is found between the NCEs (M) and the outcomes of interest (Y), or between the NCOs (W) and the target treatments/exposures (X), this suggests the presence of confounding.

In 2013, Groenwold proposed the concept of negative control time, positing that variations in the treatment/exposure-outcome associations observed in different study periods may suggest the presence of confounding.⁵⁶ Further, the method of negative control population was informally adopted across medical researches.⁵⁷ Such analyses consider a group comparable to the study population but with established null treatment/exposure effects on the outcome through causal pathways. In recent years, other NC methods like the double negative control (DNC), the control outcome calibration approach (COCA), the single proxy control, and the empirical calibration have developed rapidly.⁵⁸⁻⁶¹ By integrating NC designs, researchers can enhance the robustness of effect estimation and mitigate biases arising from confounders, marking a significant advancement in causal inference methodology. Further, the NC methods can be used in combination with the DID method to detect bias in treatment/exposure effect estimates, particularly under weaker assumptions.^{62,63}

Model specification. Taking the use of NCOs as an example, this approach extends the DID method into an NCO framework to adjust for bias arising from unmeasured confounding in linear models. By leveraging the prior assumption that "the target treatments/exposures (X) do not causally affect NCOs (W)", any observed association between the treatments/exposures (X) and NCOs (W) is attributed to the effect of unmeasured confounders (U). When constructing a linear model of $W \sim X$, a statistically significant association suggests the presence of unmeasured confounders (U) and infers its confounding effects on treatments/exposures (X) and outcomes (Y).⁵⁸ If the unmeasured confounders (U) were measured, the outcome variables (Y) could be regressed on the treatment/exposure variables (X) and unmeasured confounders (U). This would allow fitting the correct linear model of $Y \sim X + U$, from which the true causal effects of the treatments/exposures (X) on the outcomes (Y) could be obtained. However, since U is unmeasured in reality, the fitted model is obtained by using a proxy for U based on X , for example, via the model of $U \sim X$. Consequently, the estimated coefficient of X in this model combines both the true causal effects of X on Y and the effect of U . Critically, the confounding effects cannot be directly separated from the true causal effects because U is not observed. Conversely, in the NCOs model of $W \sim X$, since there is no causal effect of X on W , the coefficient of X quantifies solely the magnitude of confounding bias.⁵⁸

The selection of NCOs. Building on the core principle of NCOs, the empirical calibration method leverages a set of NCOs (W) to address both random error and systematic bias in observational studies, thereby refining causal inference for the X - Y association.⁶⁴ The selection of NCOs (W) is not random, must meet three key criteria, which directly determine the reliability of bias adjustment. First, there must be no direct or indirect causal association between the NCOs (W) and the target exposures (X). For example, when studying the effects of smoking (X) on lung cancer (Y), bone mineral density (W) can serve as a candidate NCO, as there is no clear clinical evidence supporting a mechanism by which smoking directly or indirectly affects bone mineral density. Second, the NCOs (W) and the outcomes of interest (Y) must be influenced by the same unmeasured confounders. For example, unmeasured confounders such as socioeconomic status may affect both smoking behavior and lung cancer risk. If the selected NCO is also influenced by socioeconomic status, it can capture bias consistent with that in the X - Y

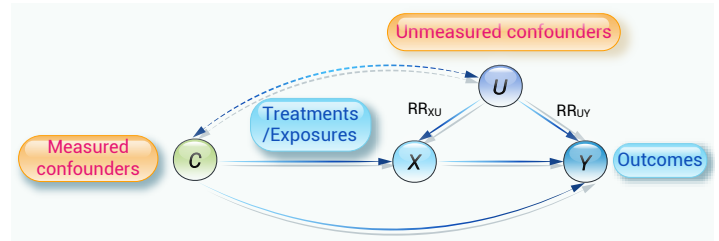


Figure 6. Causal diagram (DAG) of confounding structure RR_{XU} denotes the risk ratio (RR) of unmeasured confounders (U) with respect to treatments/exposures (X), and RR_{UY} denotes the RR of unmeasured confounders (U) with respect to outcomes (Y). The E-value is the minimum strength of association (on the RR scale) that unmeasured confounders (U) would need to have with both the exposures (X) and the outcomes (Y) to fully explain away the observed X - Y association.

association. Third, there should be no direct causal relationship between the NCOs (W) and the outcomes of interest (Y). For example, chronic obstructive pulmonary disease (COPD) cannot be used as an NCO in smoking-lung cancer studies, as clinical evidence confirms that COPD itself increases lung cancer risk. Using it as an NCO would lead to interference from the W - Y causal association in the X - W association, distorting bias estimation.⁵⁵ In practical application, the R package *EmpiricalCalibration* can help quickly screen candidate NCOs that meet the above criteria, and semi-automated predictive models can further optimize NCO selection.^{61,65} These tools reduce subjective bias in the selection process, ensuring that the final selected NCOs effectively support bias adjustment and make the causal inference of X - Y association more aligned with clinical reality.

Case study. Lane et al. conducted a new-user cohort analysis among patients with rheumatoid arthritis aged 18 years or older. This analysis compared those initiating treatment with hydroxychloroquine versus sulfasalazine for the risk of depression, suicidal ideation and mental illnesses, with a 30-day follow-up period.⁶⁶ Data for the study were derived from 14 databases across different countries, including claims data and EHRs. To assess the potential interference of unmeasured confounders on the main findings, sensitivity analysis using NCOs was conducted. A set of 65 NCOs, such as hypertrophic scar, myoclonus and snoring, having no causal relationship with either hydroxychloroquine or sulfasalazine, were identified using a semi-automated selection process based on a predictive model. Empirical calibration method was used to quantify the distribution of potential bias, thereby minimizing the effects of unmeasured confounders on the estimates. The study results showed that the HRs for almost all NCOs were approximately 1, indicating that the risk of the study conclusions being affected by unmeasured confounders is minimal.

SENSITIVITY ANALYSES FOR THE IMPACT OF UNMEASURED CONFOUNDERS

E-value

The E-value was first introduced as a standardized sensitivity analysis tool by Ding & VanderWeele in 2017.⁶⁷ It provides a quantitative measure of the minimum association strength that the unmeasured confounders (U) should have with both the treatments/exposures (X) and the outcomes (Y) to entirely explain away the observed treatment/exposure-outcome associations, after adjusting for all measured confounders (C). Fundamentally, the E-Value serves as a threshold analysis method that does not directly estimate the magnitude of unmeasured confounding bias, but calculates the minimum strength of such bias needed to reverse the current conclusion (Figure 6).⁶⁸

Model specification. Unlike traditional statistical methods for dealing with confounders, the E-value requires no assumption about the distributions of confounders or the specifics of causal mechanisms, thereby enabling direct assessments of result robustness against unmeasured bias.⁶⁹ The E-value derives solely from the observed associations between the target treatments/exposures and the outcomes of interest. Specifically, for a given risk ratio (RR) that quantifies this association, the E-value is calculated as the RR plus the square root of quantity RR multiplied by (RR minus 1).⁷⁰ This approach can be adapted to odds ratio (OR) and hazard ratio (HR) when the outcome is rare. However, for the common outcomes, the OR or HR must be

modified, by replacing RR with the square root of the OR or HR.⁷¹ A large E-value indicates that considerable unmeasured confounders would be needed to explain away the observed treatment/exposure-outcome association. Conversely, a small E-value suggests that even weak unmeasured confounders could suffice to explain this association. Thus, lower E-values make unmeasured confounding a more plausible explanation than higher E-values.⁶⁷

Limitations. However, the E-value has inherent limitations. It assumes a symmetric relationship between unmeasured confounders and both the treatments/exposures and outcomes, which is often inconsistent with real-world scenarios. Moreover, the E-value is inherently susceptible to these embedded biases in the original study and monotonically related to the effect estimators (as RR). If such biases exist, the E-value becomes distorted, failing to accurately reflect the true impact of unmeasured confounders. In essence, the E-value serves solely as a sensitivity analysis tool and cannot substitute for the rigorous research design indispensable to observational studies.⁷²

Case study. David et al. explored the associations between bariatric surgery and the occurrence of macrovascular diseases in patients with type 2 diabetes and severe obesity.⁷³ The findings revealed that bariatric surgery was associated with a lower incidence of 5-year compound macrovascular events, with a HR of 0.60 (95%CI: 0.42-0.86). Sensitivity analysis using the E-value method was conducted to estimate what the relative risk would have to be for any unmeasured confounder to overcome the observed associations of bariatric surgery with macrovascular disease in this study. An E-value of 2.72 strengthened the evidence that bariatric surgery reduced macrovascular disease risk in patients with type 2 diabetes and severe obesity. It confirmed that the observed 40% risk reduction (HR = 0.60) was not easily explained by unmeasured confounders and the true associations were likely to be real.

Other methods

Sensitivity analysis methods for unmeasured confounders vary across the research question and data type. In 2017, Kasza et al. proposed the method of confounding function, which is based on the counterfactual theory.⁷⁴ This method quantifies the magnitude of unmeasured confounding by constructing a confounding function. In 2016, Ding and VanderWeele proposed the method of bounding factor.⁶⁹ This approach requires no assumptions about unmeasured confounders and can assess the influence of all confounders.

In practice, researchers may simulate unmeasured confounding under various scenarios using parameters like the proportions of unmeasured confounders, the correlation coefficients of treatments/exposures and outcome with unmeasured confounders. If sensitivity analysis reveals substantial confounding effects on outcome estimation, observed treatment/exposure-outcome associations may be invalidated. This approach imposes no distributional restrictions on unmeasured confounders.⁷⁵⁻⁷⁷

External database-informed methods represent a recent research frontier. When external datasets contain richer information than the primary study, unmeasured confounders in the primary study data may be considered as missing values and imputed using techniques like multiple imputation or Bayesian imputation. Beyond imputation, sensitivity analysis based on external information borrowing include Monte Carlo sensitivity analysis (MCSA) and Bayesian sensitivity analysis (BSA).⁷⁸ MCSA assumes the distribution of unmeasured confounders from external data to simulate data generation to handle unmeasured confounding. BSA incorporates prior distribution from external data with simulated confounder samples to derive posterior distribution for parameter estimation.

STATISTICAL METHODS FOR DEALING WITH TIME-VARYING CONFOUNDERS

Marginal structural model (MSM)

Marginal structural model (MSM) was first proposed by Robins in 1998.⁷⁹ This approach was developed to analyze the longitudinal data involving time-varying treatments/exposures and time-varying confounders. Time-varying confounders generally feature three characteristics: (1) changing over time; (2) influencing outcomes; (3) being affected by prior treatments/exposures while influencing subsequent treatments/exposures.⁸⁰ Rooted in the counter-

factual framework, MSM posits that each individual has potential outcomes under all possible treatment/exposure scenarios ("treated/exposed" vs. "untreated/unexposed"). The causal effect is defined as the difference in the probability distribution of these potential outcomes. However, only one outcome per individual is observed in practice, making direct counterfactual measurement impossible.⁸¹

To obtain unbiased effect estimates, MSM employs IPTW to weight the original population, by creating a pseudo-population where the treatment/exposure history is independent of measured confounders.⁸² By adjusting for the influence of both time-varying and baseline confounders through this weighting scheme, MSM enables unbiased estimation of the causal associations between the treatments/exposures and the outcomes. This approach effectively mimics a randomized trial setting in observational studies, ensuring that the estimated treatment/exposure effects are free from confounding bias.

Model specification. The MSM specifies the relationship between the marginal mean of potential outcomes (Y^{a,x_1}) and the treatment/exposure history (x_0, x_1), where x_0 denotes the baseline exposure, x_1 denotes the follow-up exposure and Y^{a,x_1} denotes the potential outcome that would be observed if the exposures were set to specific levels. This approach estimates the average causal effects at the population level after adjustments for all sources of confounders. Specifically, consider a parameterized model that includes an intercept term (β_0) for baseline outcome under no treatment/exposure at both time point, two separate effect terms for each treatment/exposure time point (x_0, x_1), and an interaction effect term for concurrent treatments/exposures (x_0, x_1). The primary interest lies in comparing the expected potential outcomes between outcomes under consistent treatment/exposure ($x_0=1, x_1=1$) versus consistent non-treatment/non-exposure ($x_0=0, x_1=0$). This population-averaged effect combines the individual treatment/exposure effects from each time period along with their joint effects.⁸¹

Case study. Barbulescu et al. explored the causal associations between glucocorticoid use and the risk of serious infections in patients with rheumatoid arthritis.⁸³ This study used MSM primarily to address time-varying confounders, which is a key limitation of traditional analytical methods cannot overcome. In rheumatoid arthritis care, glucocorticoid use is dynamic that clinicians adjust dosing, initiate or discontinue it based on the real-time conditions of patients. Meanwhile, confounders like current disease activities and concomitant medications also change over time. Disease activities are shaped by prior glucocorticoid use and guide current prescribing, while concomitant medications are modified with glucocorticoid regimens. By constructing time-varying IPTWs, MSM creates a pseudo-population where confounders are balanced across groups, mimicking the balance in RCTs.

SOFTWARE IMPLEMENTATION

In observational CER, addressing confounders relies heavily on specialized software tools. R, SAS, and Stata offer comprehensive packages, procedures and commands to implement various statistical methods. Table 2 summarizes the key tools in these software packages for different methods.

CRITICAL ROLE OF SENSITIVITY ANALYSES IN CONFOUNDER HANDLING

In the observational CER, confounder handling is a challenge that spans the entire research process, from design to analysis. The complexity of confounding structures makes it difficult for any single analytical method to eliminate all bias. This inherent uncertainty underscores the paramount importance of conducting comprehensive sensitivity analyses. By systematically evaluating the robustness of study findings to potential violations of key assumptions, sensitivity analyses help quantify the uncertainty of their conclusions and assess the strength of evidence supporting causal relationships.⁸⁴⁻⁸⁵

Sensitivity analysis strategies in the study design phase

Sensitivity analyses during the study design phase aim to evaluate the impact of uncertainties in key definitions and design choices on the final conclusions, typically involving modifications of study definitions and modifi-

Table 2. Tool summary of R/SAS/Stata.

Category	Methods	R packages ^a	SAS procedures ^a	Stata commands ^a
Measured confounders	Restriction	dplyr (v3.5.0+)	DATA STEP (base) PROC STEP (base)	keep (base), drop (base)
	Matching	MatchIt (v3.6.0+), Matching (v2.6.0+), optmatch (v3.5.0+)	PROC PSMATCH (SAS/STAT 14.2+), PROC SURVEYSELECT (SAS/STAT (R) 9.2+)	psmatch2 (v11+) ^b , teffects nnmatch (v13+)
	Stratification	survival (v3.5.0+), survey (v4.1.0+)	PROC SURVEYFREQ (SAS/STAT (R) 9.2+), PROC LOGISTIC (SAS/STAT (R) 9.2+)	stssplit (v7+), stratif (v7+)
	Standardization	stdReg2 (v4.1.0+), epiR (v3.0.0+), sjPlot (v4.1.0+)	PROC STD RATE (SAS/STAT (R) 12.3+)	stdize (v6+), standsurv (v14+) ^b
	Regression Adjustment	stats (base), survival (v3.5.0+), rms (v4.4.0+)	PROC REG (SAS/STAT (R) 9.2+), PROC GLM (SAS/STAT (R) 9.2+), PROC LOGISTIC (SAS/STAT (R) 9.2+), PROC PHREG (SAS/STAT (R) 9.2+)	regress (base), logistic (base), stcox (v7+)
	Propensity score matching (PSM)	Matching (v2.6.0+), MatchIt (v3.6.0+), PSweight (v3.5.0+), WeightIt (v4.1.0+), twang (v2.10+)	PROC PSMATCH (SAS/STAT 14.2+)	pscore (v8+), psmatch2 (v11+) ^b , teffects (v13+)
	Inverse probability treatment weighting (IPTW)	WeightIt (v4.1.0+), ipw (v3.0.0+), twang (v2.10+), PSW (v3.0.0+)	PROC GENMOD (SAS/STAT (R) 9.2+)	ipwpoint (v11+) ^b , teffects (v13+), ipw (v11+) ^b
	PS stratification or subclassification	MatchIt (v3.6.0+), PSweight (v3.5.0+)	PROC PSMATCH (SAS/STAT 14.2+)	psmatch2 (v11+) ^b , teffects (v13+)
	Covariate adjustment using the PS	<i>Include PS as covariate in regression models</i>	<i>Include PS as covariate in regression models</i>	pscore (v8+), <i>include PS as covariate in regression models</i>
	Disease risk score (DRS)	riskRegression (v3.5.0+), rms (v4.4.0+)	PROC LOGISTIC (SAS/STAT (R) 9.2+), PROC PHREG (SAS/STAT (R) 9.2+)	drscore (v14+) ^b , teffects (v13+)
Unmeasured confounders	Instrumental variable (IV)	AER (v3.0.0+), ivreg (v3.6.0+), fixest (v3.5.0+), lfe (v2.15.2+)	PROC QLIM (SAS/ETS (R) 9.2+), PROC SYSLIN (SAS/ETS (R) 9.2+)	ivregress (v10+), ivreg2 (v9+) ^b , xtivreg (v9+)
	Difference-in-differences (DID)	fixest (v3.5.0+), plm (v3.2.0+), did (v3.5.0+)	PROC PANEL (SAS/ETS (R) 9.2+), PROC GENMOD (SAS/STAT (R) 9.2+)	xtreg (v8+), diff (v11+), areg (v6+)
	Negative control (NC)	NegativeControlOutcomeAdjustment (v3.5.0+), EmpiricalCalibration (v3.5.0+)	<i>Manual</i>	rcr (v12+) ^b
	Prior event rate ratio adjustment (PERR)	tidyverse (v3.3.0+), boot (v3.0.0+)	PROC GENMOD (SAS/STAT (R) 9.2+)	gen (base)
	Regression discontinuity design (RDD)	rdd (v3.0.0+), rdrobust (v3.1.1+), rddensity (v3.1.0+)	PROC LOESS (SAS/STAT (R) 9.2+), PROC NLIN (SAS/STAT (R) 9.2+)	rdrobust (v13+) ^b , rddensity (v13+) ^b
Sensitive analyses for unmeasured confounders	E-value	EValue (v3.5.0+)	PROC CAUSALMED (SAS/STAT 14.3+)	evaluate (v14+) ^b
	Confounding function	tipr (v2.10+), sensemakr (v3.1.0+)	<i>Manual</i>	conbounds (v15+) ^b
	Bounding factor	tipr (v2.10+), sensemakr (v3.1.0+)	<i>Manual</i>	ebounds (v14+) ^b
	Monte Carlo sensitivity analysis (MCSA)	sensemakr (v3.1.0+)	PROC MCMC (SAS/STAT (R) 9.2+)	mcsensitivity (v14+) ^b
	Bayesian sensitivity analysis (BSA)	rstan (v3.4.0+), brms (v3.6.0+), BayesSenMC (v3.4.0+)	PROC MCMC (SAS/STAT (R) 9.2+)	bayesmh (v14+)
Time-varying confounders	Marginal structural model (MSM)	ipw (v3.0.0+), WeightIt (v4.1.0+), tmle (v3.6.0+)	PROC GENMOD (SAS/STAT (R) 9.2+), PROC PHREG (SAS/STAT (R) 9.2+)	ipw (v11+) ^b , teffects ipw (v13+)
	Structural nested model (SNM)	gesttools (v3.5.0+), gfoRmula (v3.5.0+)	PROC IML (SAS/IML (R) 12.3+), PROC NL MIXED (SAS/STAT (R) 9.2+), PROC SYSLIN (SAS/ETS (R) 9.2+)	snm (v13+) ^b

Note: ^a The contents within the parentheses are the software version numbers; ^b An external command, must be installed via Statistical Software Components (SSC) repository.

cations of the study design.

Modifying study definitions requires employing different algorithms or criteria to specify key study components. Exposure might be defined according to stricter medication adherence criteria (e.g., requiring at least four prescriptions rather than just the first prescription). Outcome might be

defined according to standard definitions such as ICD codes or refined to more clinically specific diagnostic events (e.g., diagnosing nonsteroidal anti-inflammatory drug related myocardial infarction requiring simultaneous records of elevated myocardial enzymes). Study population might be defined using different inclusion and exclusion criteria (e.g., such as adjusting mini-

mum follow-up time requirements or excluding patients with specific comorbidities to test for selection bias).

Modifying study design requires structural adjustments to the research framework to evaluate the stability of findings. Common methods include altering the study time window, (e.g., moving the baseline period in policy impact studies) and replacing the control group (e.g., changing from all non-treated/exposed individuals to PSM controls). Moreover, it involves substituting between different real-world data sources to examine database-specific effects, and integrating multiple complementary data sources, such as combining EHR with pharmacy data to enhance representativeness and reduce single-source bias.

Sensitivity analysis strategies in the data analysis phase

Sensitivity analyses during the data analysis phase focus on evaluating how statistical models and methodological choices might influence effect estimates. Common strategies include adjusting model specifications, and conducting cross-validations with different analytical methods. Adjusting model specifications examines whether the results are sensitive to specific modeling assumptions, thereby reducing the risk of bias from over-adjustment or under-adjustment. Typical implementations involve varying the combination of covariates (e.g., comparing the basic model with only age and gender to an extended model that includes additional clinical indicators), and applying different regression model types (e.g., comparing logistic regression model with Poisson regression model). Conducting cross-validations with different analytical methods reduces potential biases that may arise from relying exclusively on a single method, improving the reliability and persuasiveness of the conclusions. These implementations include repeating effect estimation using different confounder handling methods (e.g., comparing the results of PSM and DRS stratification) and employing alternative methods for handling missing data (e.g., replacing complete case analyses with multiple imputation methods).

Analysis targeting unmeasured confounders provides essential assessment of potential bias when direct measurement and adjustment of all confounders proves impossible. Quantitative methods like E-values offer intuitive measurements by quantifying how strongly an unmeasured confounder would need to be associated with both the exposures/treatments and the outcomes to fully explain away the observed effects. Simulation-based approaches, including MCSA and BSA, allow researchers to incorporate external knowledge or prior distributions to simulate how unmeasured confounding might impact effect estimates under various hypothetical scenarios. When multiple methodologically independent approaches yield effect estimates that are consistent in both direction and magnitude, the credibility of study conclusions becomes substantially strengthened.

Ultimately, while any observational study is inherently limited in its ability to definitively establish causality in the absence of randomization, a well-designed and rigorously executed sensitivity analysis framework significantly enhances the credibility of research findings. Such a framework transparently communicates the degree of uncertainty to readers, clearly defines the boundary conditions under which conclusions remain valid, and demonstrates a commitment to rigorous self-scrutiny. Therefore, investing in comprehensive sensitivity analyses is indispensable for generating reliable and interpretable evidence from observational CER, ultimately supporting more informed clinical and policy decisions.

REPORTING STANDARDS AND WRITING GUIDELINES

In observational CER, clearly defined writing guidelines and reporting standards are fundamental to ensure the transparency, reproducibility and credibility. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement offers a robust framework for reporting observational studies.⁸⁶ Additionally, the REporting of studies Conducted using Observational Routinely collected health Data (RECORD) statement and its extension for pharmacoepidemiology (RECORD-PE) provide further guidance tailored to routinely collected data,^{87–88} and the STROBE-MR statement and checklists offer specific reporting standards for studies employing PS and IV methods.^{89–91} Building on these foundational standards, we propose the following supplementary checklists and writing guidelines for reporting the identification and handling of confounders in observational CER. Authors

should explicitly address each of the following items in the relevant sections of their manuscripts (Table 3).

Title & Abstract

The title should clearly identify the study as an observational CER and specify the specified treatments/exposures under comparison and the primary outcomes, e.g., "Comparative Effectiveness of Intravascular Microaxial Left Ventricular Assist Device versus Intra-aortic Balloon Pump on In-hospital Mortality among Patients with Acute Myocardial Infarction: An Observational Study". In the abstract, authors should briefly summarize the most predominant confounders and key methods for addressing them, enabling readers to quickly grasp the core strategy for addressing confounding bias. The abstract should thereby provide a transparent overview that accurately represents the methodological framework supporting the causal inference of this study.

Introduction

Following the clarification of clinical significance and evidence gaps, the introduction should thoroughly elaborate on the rationale for choosing an observational design, particularly highlighting the value of observational studies in reflecting real-world clinical practice when RCTs are infeasible or unethical. Meanwhile, it should clearly demonstrate that despite non-random treatment/exposure assignments, rigorous handling of confounders in this study will significantly enhance the reliability of causal inference in this observational context, establishing this as the methodological foundation.

Methods—Study design & Participants

This section should provide a comprehensive description of study design, data source, sampling strategy and inclusion/exclusion criteria. It is essential to explain how each of these elements is selected to minimize selection bias and handle for potential confounding. Authors should clearly specify the study design employed (e.g., cohort design, case-control design, new-user design, active comparator design), and provide a detailed rationale for its selection. For example, describing how a new-user design helps prevalent user bias by including only participants who are initiating treatments/exposures, or how an active comparator design enhances comparability by restricting the study population to patients with similar clinical indications. A thorough description of the data source is essential, including the specific type of data utilized (e.g., EHRs, administrative claims databases, disease registries), the population covered, the geographical and temporal scope, and the data collection processes. The sampling strategy should be clearly specified, whether it involves the entire eligible population or a selected subset through methods such as matching or stratification. The justification for this approach should explain how it contributes to achieving covariate balance between comparison groups or improves the efficiency of the analysis. Furthermore, each inclusion and exclusion criterion should be presented with a clear scientific rationale, explaining how it helps create a more comparable study population and reduce initial confounding.

Methods—Variable definitions

Precise operational definitions for all key variables, including treatments/exposures, outcomes, and predominant confounders, are fundamental to establishing a transparent and methodologically sound foundation for causal inference. The timing of measurement for key variables, particularly the treatments/exposures, must be explicitly defined, as it is critical for establishing temporal precedence and correctly classifying other variables (e.g., confounders or mediators). The process of identifying and specifying confounders should be systematically justified, detailing whether the selection was informed by clinical expertise, systematic literature reviews, or causal diagrams such as DAGs.⁹² The use of DAGs is highly recommended to visually represent assumed causal pathways and guide the identification of a sufficient adjustment set, thereby preventing inappropriate adjustment for mediators or colliders. For each confounder, an explicit operational definition must be provided, including its measurement method (e.g., specific diagnostic codes, laboratory thresholds, or validated instruments), data source (e.g., EHRs, administrative claims databases, disease registries), and a clearly defined assessment time window that logically precedes exposure initiation to ensure temporal precedence.

Table 3. Supplementary checklist for reporting confounder handling in observational CER.

Manuscript section	Core reporting elements	Specific requirements for confounder handling (Checklist)
Title & Abstract	Clarify study type, comparisons, primary outcomes, and key methods	■ Does the title identify the study as an "observational comparative effectiveness study"? ■ Does the title specify the compared treatments/exposures and primary outcomes? ■ Does the abstract summarize the predominant confounders and key methods used to address them?
Introduction	Elucidate research necessity and rationale for observational design	■ Does it elaborate the rationale for choosing an observational design? ■ Does it demonstrate that rigorous handling of confounders is key to enhancing the reliability of causal inference in this observational context?
Methods — Study design & Participants	Describe study design, data source, sampling strategy, and study population	■ Does it explain how the study design helps reduce selection bias and potential confounding? ■ Does it describe the data source, including the specific type of data utilized, population coverage, geographical and temporal scope, and data collection processes? ■ Does it specify how the sampling strategy to achieve covariate balance or improve analytical efficiency? ■ Does it state how each inclusion and exclusion criterion explain to create a more comparable study population and reduce initial confounding?
Methods — Variable definitions	Clearly define treatments/exposures outcomes, and confounders	■ Does it list candidate confounders with their selection rationale? ■ Does it define the measurement method, data source, and assessment time window for each confounder?
Methods — Statistical analyses	Detail statistical methods for confounder handling	■ Method Justifications: Does it provide the rationale for choosing specific methods? ■ Model specification Details: - PS methods: Does it list all covariates included in the PS model, and report the specific model type and matching/weighting parameters? - IV methods: Does it discuss the validity of the IVs? - DID methods: Does it state and empirically assess the parallel trends assumption for DID? - NC methods: Does it specify whether a NCE or NCO was used and provide the rationale for its selection? - E-value method: Does it specify the effect estimate used for the E-value calculation? - MSM: Does it describe the calculation of time-varying weights for MSM? ■ Missing data handling: Does it report the extent and handling method of missing data for key confounders? ■ Sensitivity analysis plan: Does it describe plans to assess unmeasured confounders using methods like E-value or NC?
Results	Present covariate balance assessment, effect estimate presentation, and sensitivity analysis results	■ Covariate balance assessment: Does it quantitatively report the balance of covariates between groups after using PS methods? ■ Effect estimate presentation: Does it present both unadjusted and confounder-adjusted effect estimates with confidence intervals? ■ Sensitivity analysis results: Does it report the results of sensitivity analyses to assess robustness to unmeasured confounding?
Discussions	Interpret results, compare with literature, and acknowledge limitations	■ Literature comparison: Does it discuss differences in confounder handling methods when comparing with previous studies? ■ Limitations: Is residual confounding explicitly discussed? - Does it explicitly discuss residual confounding? - Does it identify specific potentially important unmeasured or imperfectly measured confounders? - Does it consider the potential direction of bias from these residual confounders? - Does it critically reflect upon the core assumptions of the primary statistical methods?

Methods—Statistical analyses

Method justifications. The statistical methods for confounder handling should be explicitly specified based on the nature of the confounders. Specifically, for measured confounders, appropriate methods include multivariate regression, PS and DRS. For unmeasured confounders, appropriate methods include IV, DID, NC and E-value. In studies involving time-varying confounders affected by prior treatments/exposures, MSM represents the recommended method. For each method selected, authors must provide a clear rationale that aligns with the study objectives, data structure, and the nature of the treatment/exposure-outcome relationships. For example, stating that "IPTW was selected to preserve the full sample size and enhance generalizability", "DRS was chosen due to the rarity of the exposure", "IV method was employed to address potential unmeasured confounding", "Negative control outcomes were utilized to test for residual bias".

Method specification details. A detailed description of all statistical methods is necessary to ensure reproducibility. For PS or DRS methods, this

includes listing all covariates included in the PS or DRS estimation, specifying the model type (e.g., logistic regression), reporting parameters such as matching ratio, caliper width, and algorithm for matching, or describing weight calculation and truncation criteria for IPTW. For IV method, the validity of the IVs must be thoroughly discussed. For DID method, the parallel trends assumption must be explicitly stated and empirically tested. For MSM, the process for calculating time-varying inverse probability weights must be explicitly described, including how time-dependent confounders were addressed in the weight estimation.

Missing data handling. The handling of missing data in key confounders must be clearly described, including the proportion of missing values and the specific approach used (e.g., complete-case analysis, multiple imputation).

Sensitivity analysis plan. A pre-specified sensitivity analysis framework should be established to evaluate the robustness of primary findings to potential violations of key assumptions. This must include quantitative assessments using E-values to determine the strength of unmeasured

confounding required to nullify the observed associations, as well as NC analyses employing either NCEs or NCOs to detect residual confounding. Additional sensitivity analyses should explore the impact of alternative model specifications, such as varying the set of covariates included in PS models, using different matching algorithms, or applying alternative approaches to handling missing data. For IV methods, sensitivity tests examining the consequences of potential violations of the exclusion restriction should be incorporated. All sensitivity analyses should be clearly motivated by specific methodological concerns relevant to the study context, with results interpreted in terms of their implications for the validity of the primary conclusions.

Results

Covariate balance assessments. Following the application of PS methods, the balance of covariates between exposure groups must be quantitatively reported. Authors should present standardized mean differences (SMDs) for all key confounders, both before and after adjustment, with SMD values below 0.1 generally indicating adequate balance. The use of visual tools such as Love plots is strongly recommended to provide an intuitive display of covariate balance across all measured variables. This comprehensive reporting enables readers to critically evaluate the success of the confounder adjustment strategies employed.

Effect estimate presentations. The results should include a complete presentation of effect estimates at multiple levels of adjustment. Specifically, authors should report both unadjusted estimates and confounder-adjusted estimates (e.g., HR, OR) alongside their corresponding confidence intervals. It is essential to present estimates from different adjusted models, including demographics-only adjustments and fully adjusted models that incorporate all identified confounders through the primary statistical methods.

Sensitivity analysis results. This section should report the specific results of all pre-specified sensitivity analyses. Authors should present the results of E-values for both the point estimate and the confidence interval. Results from NC methods should be presented to empirically evaluate the potential presence of residual confounding. In addition to the primary analysis method (e.g., PSM), other alternative methods (e.g. DRS) should also be applied in sensitivity analyses, and the results from these alternative methods should be reported. The interpretation of these sensitivity analyses should directly address their implications for the robustness of the primary findings and the plausibility of causal interpretations given potential unmeasured confounding.

Discussions

When interpreting the study findings, the discussion should systematically address methodological considerations related to confounder handling. In comparing results with previous literature, authors should explicitly discuss differences in confounder handling methods across studies and analyze how these methodological variations might explain difference in observed effects. The discussion must include a dedicated section addressing limitations, with particular focus on residual confounding. This should involve specifying important unmeasured or imperfectly measured confounders, such as genetic factors, health behaviors, or socioeconomic details, discussing the likely direction of bias these might introduce (overestimation or underestimation of true effects), and critically reflecting on potential challenges to the core assumptions of the primary statistical methods used in the study context.

CONCLUSIONS AND PERSPECTIVES

In summary, observational CER provide crucial real-world insights into the evaluation of treatment/exposure efficacy, but confounders threaten the validity of causal conclusions. This study has comprehensively reviewed statistical methods for addressing measured, unmeasured and time-varying confounders, as well as sensitivity analyses for the impact of unmeasured confounders. Among the methods reviewed, traditional approaches like restriction and regression are widely used but have inherent limitations. By contrast, advanced methods such as PS and DRS offer greater flexibility in addressing these challenges, though they require careful application. Notably, methods for unmeasured confounders often rely on, stringent assumptions, such as the independence of IVs from confounders, and the presence of parallel assumption when using DID. Assumption violations cause bias, while sensitivity analyses play a critical complementary role. They help assess the

robustness of study results by quantifying how potential residual confounding from unmeasured factors might alter causal conclusions.

Future research should focus on better addressing complex and unmeasured confounders, integrating complementary adjustment methods, and promoting standardized confounder handling practices. In parallel, it also emphasizes that integrating causal inference frameworks with emerging analytical techniques is critical to improve the validity and interpretability of real-world evidence derived from observational CER.⁹³ Strengthening the synergy between causal reasoning and data-driven approaches will further bridge methodological gaps and foster more transparent and reliable findings. Ultimately, effectively identifying and addressing confounders is vital for reliable CER, enabling more accurate causal inferences to inform clinical and public health decisions.

REFERENCES

1. Howards P. P. (2018). An overview of confounding. Part 1: the concept and how to address it. *Acta Obstet. Gynecol. Scand.* **97**:394-399. DOI:10.1111/aogs.13295
2. Hariton E. and Locascio J. J. (2018). Randomised controlled trials – the gold standard for effectiveness research. *BJOG* **125**:1716. DOI:10.1111/1471-0528.15199
3. Norris S. L., Atkins D., Bruening W., et al. (2011). Observational studies in systematic reviews of comparative effectiveness: AHRQ and the Effective Health Care Program. *J. Clin. Epidemiol.* **64**:1178. DOI:10.1016/j.jclinepi.2010.04.027
4. Cuello-Garcia C. A., Santesso N., Morgan R. L., et al. (2022). GRADE guidance 24 optimizing the integration of randomized and non-randomized studies of interventions in evidence syntheses and health guidelines. *J. Clin. Epidemiol.* **142**:200. DOI:10.1016/j.jclinepi.2021.11.026
5. Meuli L. and Dick F. (2018). Understanding confounding in observational studies. *Eur. J. Vasc. Endovasc. Surg.* **55**:737. DOI:10.1016/j.ejvs.2018.02.028
6. Kyriacou D. N. and Lewis R. J. (2016). Confounding by indication in clinical research. *JAMA* **316**:1818. DOI:10.1001/jama.2016.16435
7. Prada-Ramallal G., Takkouche B. and Figueiras A. (2019). Bias in pharmacoepidemiologic studies using secondary health care databases: A scoping review. *BMC Med. Res. Methodol.* **19**:53. DOI:10.1186/s12874-019-0695-y
8. Toh S., Reichman M. E., Graham D. J., et al. (2018). Prospective postmarketing surveillance of acute myocardial infarction in new users of Saxagliptin: A population-based study. *Diabetes Care* **41**:39. DOI:10.2337/dc17-0476
9. Varga A. N., Guevara Morel A. E., Lokkerbol J., et al. (2023). Dealing with confounding in observational studies: A scoping review of methods evaluated in simulation studies with single-point exposure. *Stat. Med.* **42**:487. DOI:10.1002/sim.9628
10. D'Onofrio B. M., Sjölander A., Lahey B. B., et al. (2020). Accounting for confounding in observational studies. *Annu. Rev. Clin. Psychol.* **16**:25. DOI:10.1146/annurev-clinpsy-032816-045030
11. Guo J., Wang T., Cao H., et al. (2025). Application of methodological strategies to address unmeasured confounding in real-world vaccine safety and effectiveness study: A systematic review. *J. Clin. Epidemiol.* **181**:111737. DOI:10.1016/j.jclinepi.2025.111737
12. Wei J. C., Kuo P. and Chang R. (2025). Strategies to avoid confounders and bias in observational studies. *Int. J. Rheum. Dis.* **28**:e70076. DOI:10.1111/1756-185x.70076
13. Luo H., Zhuang F., Xie R., et al. (2024). A survey on causal inference for recommendation. *The Innovation* **5**:100590. DOI:10.1016/j.xinn.2024.100590
14. Lash T. L., VanderWeele T. J., Haneuse S., et al. (2021). Modern Epidemiology, Fourth Edition (Wolters Kluwer). <https://shop.lww.com/ModernEpidemiology/p/9781451193282?srsltid=AfmBOopZYdqNHzVrTuna07KJSNgMQ7v9W00q8Wxb6F8xBn086utVlKPF>.
15. Velentgas P., Dreyer N. A., Nourjah P., et al. (2013). Developing a protocol for observational comparative effectiveness research: A user's guide. In (Agency for Healthcare Research and Quality (US)). <https://www.ncbi.nlm.nih.gov/books/NBK126190/>.
16. Rubin D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29**:159-183. DOI:10.2307/2529684
17. Mantel N. and Haenszel W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22**:719-748.
18. Rosenbaum P. R. and Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**:41. DOI:10.1093/biomet/70.1.41
19. Andrew B. Y., Alan Brookhart M., Pearce R., et al. (2023). Propensity score methods in observational research: Brief review and guide for authors. *Br. J. Anaesth.* **131**:805-809. DOI:10.1016/j.bja.2023.06.054
20. Benedetto U., Head S. J., Angelini G. D., et al. (2018). Statistical primer: Propensity score matching and its alternatives. *Eur. J. Cardiothorac. Surg.* **53**:1112. DOI:10.1093/ejcts/ezy167
21. King G. and Nielsen R. (2019). Why propensity scores should not be used for matching. *Political Analysis* **27**:435. DOI:10.1017/pan.2019.11
22. Austin P. C. and Stuart E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* **34**:3661. DOI:10.1002/

- sim.6607
23. Desai R. J. and Franklin J. M. (2019). Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: A primer for practitioners. *BMJ* **367**:i5657. DOI:10.1136/bmj.i5657
24. Cole S. R. and Hernán M. A. (2008). Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.* **168**:656. DOI:10.1093/aje/kwn164
25. Neuhäuser M., Thielmann M. and Ruxton G. D. (2018). The number of strata in propensity score stratification for a binary outcome. *Arch. Med. Sci.* **14**:695. DOI:10.5114/aoms.2016.61813
26. Austin P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav. Res.* **46**:399. DOI:10.1080/00273171.2011.568786
27. Zou B., Zou F., Shuster J. J., et al. (2016). On variance estimate for covariate adjustment by propensity score analysis. *Stat. Med.* **35**:3537. DOI:10.1002/sim.6943
28. Dhruva S. S., Ross J. S., Mortazavi B. J., et al. (2020). Association of use of an intravascular microaxial left ventricular assist device vs intra-aortic balloon pump with in-hospital mortality and major bleeding among patients with acute myocardial infarction complicated by cardiogenic shock. *JAMA* **323**:734. DOI:10.1001/jama.2020.0254
29. Miettinen O. S. (1976). Stratification by a multivariate confounder score. *Am. J. Epidemiol.* **104**:609. DOI:10.1093/oxfordjournals.aje.a112339
30. Glynn R. J., Gagne J. J. and Schneeweiss S. (2012). Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiol. Drug Saf.* **21** Suppl 2:138. DOI:10.1002/pds.3231
31. Wyss R., Lunt M., Brookhart M. A., et al. (2014). Reducing bias amplification in the presence of unmeasured confounding through out-of-sample estimation strategies for the disease risk score. *J. Causal Inference* **2**:131. DOI:10.1515/jci-2014-0009
32. Tadrous M., Gagne J. J., Stürmer T., et al. (2013). Disease risk score as a confounder summary method: Systematic review and recommendations. *Pharmacoepidemiol. Drug Saf.* **22**:122. DOI:10.1002/pds.3377
33. Kumamaru H., Schneeweiss S., Glynn R. J., et al. (2016). Dimension reduction and shrinkage methods for high dimensional disease risk scores in historical data. *Emerg. Themes Epidemiol.* **13**:5. DOI:10.1186/s12982-016-0047-x
34. Blin P., Dureau-Pournin C., Jové J., et al. (2020). Secondary prevention of acute coronary syndrome with antiplatelet agents in real life: A high-dimensional propensity score matched cohort study in the French National claims database. *MethodsX* **7**:100796. DOI:10.1016/j.mex.2020.100796
35. Davies N. M., Smith G. D., Windmeijer F., et al. (2013). Issues in the reporting and conduct of instrumental variable studies: A systematic review. *Epidemiology* **24**:363. DOI:10.1097/EDE.0b013e31828abaf8
36. Baiocchi M., Cheng J. and Small D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33**:2297. DOI:10.1002/sim.6128
37. Guo Z., Cheng J., Lorch S. A., et al. (2014). Using an instrumental variable to test for unmeasured confounding. *Stat. Med.* **33**:3528. DOI:10.1002/sim.6227
38. Larsson S. C., Butterworth A. S. and Burgess S. (2023). Mendelian randomization for cardiovascular diseases: Principles and applications. *Eur. Heart J.* **44**:4913. DOI:10.1093/eurheartj/ehad736
39. Zheng J., Zhang Y., Rasheed H., et al. (2022). Trans-ethnic Mendelian-randomization study reveals causal relationships between cardiometabolic factors and chronic kidney disease. *Int. J. Epidemiol.* **50**:1995. DOI:10.1093/ije/dyab203
40. Gupta V., Walia G. K. and Sachdeva M. P. (2017). 'Mendelian randomization': An approach for exploring causal relations in epidemiology. *Public Health* **145**:113. DOI:10.1016/j.puhe.2016.12.033
41. Ference B. A., Holmes M. V. and Smith G. D. (2021). Using Mendelian Randomization to improve the design of randomized trials. *Cold Spring Harb. Perspect. Med.* **11**:a040980. DOI:10.1101/cshperspect.a040980
42. Brito C. and Pearl J. (2002). Generalized instrumental variables. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI2002)*:85. UAI-P-2002-PG-85-93.
43. Zander B. v. d., Textor J. and Liskiewicz M. (2015). Efficiently finding conditional instruments for causal inference. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence* pp:3243-3249.
44. Cheng D., Li J., Liu L., et al. (2022). Ancestral instrument method for causal inference without complete knowledge. *ArXiv Preprint ArXiv*:2201.03810. DOI:10.48550/arXiv.2201.03810
45. Cheng D., Li J., Liu L., et al. (2023). Discovering ancestral instrumental variables for causal inference from observational data. *IEEE Transactions on Neural Networks and Learning Systems* **35**:11542-11552. DOI:10.1109/TNNLS.2023.3262848
46. Cheng D., Xu Z., Li J., et al. (2024). Instrumental variable estimation for causal inference in longitudinal data with time-dependent latent confounders. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**:11480-11488. DOI:10.1609/aaai.v38i10.29029
47. Mollroy D. R., Shotwell M. S., Lopez M. G., et al. (2022). Oxygen administration during surgery and postoperative organ injury: Observational cohort study. *BMJ* **379**:e070941. DOI:10.1136/bmj-2022-070941
48. Ashenfelter O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics* **60**:47. DOI:10.2307/1924332
49. Tchetgen Tchetgen E. J., Park C. and Richardson D. B. (2024). Universal difference-in-differences for causal inference in epidemiology. *Epidemiology* **35**:16. DOI:10.1097/ede.0000000000001676
50. Wang G., Hamad R. and White J. S. (2024). Advances in difference-in-differences methods for policy evaluation research. *Epidemiology* **35**:628. DOI:10.1097/ede.0000000000001755
51. Zeldow B. and Hatfield L. A. (2021). Confounding and regression adjustment in difference-in-differences studies. *Health Serv. Res.* **56**:932-941. DOI:10.1111/1475-6773.13666
52. Goodman-Bacon A. (2021). Difference-in-differences with variation in treatment timing. *J. Econometrics* **225**:254. DOI:10.1016/j.jeconom.2021.03.014
53. Rensson A., Hudgens M. G., Keil A. P., et al. (2023). Identifying and estimating effects of sustained interventions under parallel trends assumptions. *Biometrics* **79**:2998. DOI:10.1111/biom.13862
54. Karter A. J., Parker M. M., Moffet H. H., et al. (2021). Association of real-time continuous glucose monitoring with glycemic control and acute metabolic events among patients with insulin-treated diabetes. *JAMA* **325**:2273. DOI:10.1001/jama.2021.6530
55. Lipsitch M., Tchetgen Tchetgen E. and Cohen T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* **21**:383. DOI:10.1097/EDE.0b013e3181d61eeb
56. Groenwold R. H. (2013). Falsification end points for observational studies. *JAMA* **309**:1769. DOI:10.1001/jama.2013.3089
57. Piccinini M. and Stensrud M. J. (2024). Using negative control populations to assess unmeasured confounding and direct effects. *Epidemiology* **35**:313. DOI:10.1097/ede.0000000000001724
58. Shi X., Miao W. and Tchetgen E. T. (2020). A selective review of negative control methods in epidemiology. *Curr. Epidemiol. Rep.* **7**:190. DOI:10.1007/s40471-020-00243-4
59. Park C., Richardson D. B. and Tchetgen Tchetgen E. J. (2024). Single proxy control. *Biometrics* **80**:ujae027. DOI:10.1093/biomet/ujae027
60. Li K., Emerman I., Cook A. J., et al. (2024). Using double negative controls to adjust for healthy user bias in a recombinant zoster vaccine safety study. *Am. J. Epidemiol.* **194**:2641-2649. DOI:10.1093/aje/kwae439
61. Schuemie M. J., Ryan P. B., DuMouchel W., et al. (2014). Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Stat. Med.* **33**:209. DOI:10.1002/sim.5925
62. Newsome S. J., Daniel R. M., Carr S. B., et al. (2022). Using negative control outcomes and difference-in-differences analysis to estimate treatment effects in an entirely treated cohort: The effect of ivacaftor in cystic fibrosis. *Am. J. Epidemiol.* **191**:505. DOI:10.1093/aje/kwab263
63. Sofer T., Richardson D. B., Colicino E., et al. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Stat. Sci.* **31**:348. DOI:10.1214/16-sts558
64. Schuemie M. J., Hripsak G., Ryan P. B., et al. (2016). Robust empirical calibration of p-values using observational data. *Stat. Med.* **35**:3883. DOI:10.1002/sim.6977
65. Voss E. A., Boyce R. D., Ryan P. B., et al. (2017). Accuracy of an automated knowledge base for identifying drug adverse reactions. *J. Biomed. Inform.* **66**:72. DOI:10.1016/j.jbi.2016.12.005
66. Lane J. C. E., Weaver J., Kostka K., et al. (2020). Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: A multinational, retrospective study. *Lancet Rheumatol.* **2**:e698. DOI:10.1016/s2665-9913(20)30276-9
67. VanderWeele T. J. and Ding P. (2017). Sensitivity analysis in observational research: Introducing the E-Value. *Ann. Intern. Med.* **167**:268. DOI:10.7326/m16-2607
68. Chung W. T. and Chung K. C. (2023). The use of the E-value for sensitivity analysis. *J. Clin. Epidemiol.* **163**:92. DOI:10.1016/j.jclinepi.2023.09.014
69. Ding P. and VanderWeele T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27**:368-377. DOI:10.1097/EDE.0000000000000457
70. Gaster T., Eggertsen C. M., Størring H., et al. (2023). Quantifying the impact of unmeasured confounding in observational studies with the E value. *BMJ Med.* **2**:e00366. DOI:10.1136/bmjmed-2022-000366
71. VanderWeele T. J. (2017). On a square-root transformation of the odds ratio for a common outcome. *Epidemiology* **28**:e58. DOI:10.1097/ede.0000000000000733
72. Ioannidis J. P. A., Tan Y. J. and Blum M. R. (2019). Limitations and misinterpretations of E-values for sensitivity analyses of observational studies. *Ann. Intern. Med.* **170**:108. DOI:10.7326/m18-2159
73. Fisher D. P., Johnson E., Haneuse S., et al. (2018). Association between bariatric surgery and macrovascular disease outcomes in patients with type 2 diabetes and severe obesity. *JAMA* **320**:1570. DOI:10.1001/jama.2018.14619
74. Kasza J., Wolfe R. and Schuster T. (2017). Assessing the impact of unmeasured confounding for binary outcomes using confounding functions. *Int. J. Epidemiol.* **46**:1303. DOI:10.1093/ije/dyx023
75. Schneeweiss S. (2006). Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol. Drug Saf.* **15**:291. DOI:10.1002/pds.1200
76. Lin N. X., Logan S. and Henley W. E. (2013). Bias and sensitivity analysis when estimating treatment effects from the cox model with omitted covariates. *Biometrics* **69**:850. DOI:10.1111/biom.12096

77. Groenwold R. H., Nelson D. B., Nichol K. L., et al. (2010). Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. *Int. J. Epidemiol.* **39**:107. DOI:10.1093/ije/dyp332
78. McCandless L. C. and Gustafson P. (2017). A comparison of Bayesian and Monte Carlo sensitivity analysis for unmeasured confounding. *Stat. Med.* **36**:2887. DOI:10.1002/sim.7298
79. Robins J. M., Hernán M. A. and Brumback B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**:550. DOI:10.1097/00001648-200009000-00011
80. Mansournia M. A., Etminan M., Danaei G., et al. (2017). Handling time varying confounding in observational research. *BMJ* **359**:j4587. DOI:10.1136/bmj.j4587
81. Naimi A. I., Cole S. R. and Kennedy E. H. (2017). An introduction to g methods. *Int. J. Epidemiol.* **46**:756. DOI:10.1093/ije/dyw323
82. Clare P. J., Dobbins T. A. and Mattick R. P. (2019). Causal models adjusting for time-varying confounding-a systematic review of the literature. *Int. J. Epidemiol.* **48**:254. DOI:10.1093/ije/dyy218
83. Barbulescu A., Sjölander A., Delcoigne B., et al. (2023). Glucocorticoid exposure and the risk of serious infections in rheumatoid arthritis: A marginal structural model application. *Rheumatology* **62**:3391. DOI:10.1093/rheumatology/kead083
84. Latour C. D., Delgado M., Su I. H., et al. (2025). Use of sensitivity analyses to assess uncontrolled confounding from unmeasured variables in observational, active comparator pharmacoepidemiologic studies: A systematic review. *Am. J. Epidemiol.* **194**:524. DOI:10.1093/aje/kwae234
85. Xu J., Wang Y., He Q., et al. (2025). Evaluating the agreement between sensitivity and primary analyses in observational studies using routinely collected healthcare data: A meta-epidemiology study. *BMC Med.* **23**:393. DOI:10.1186/s12916-025-04199-4
86. Von Elm E., Altman D. G., Egger M., et al. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Lancet* **370**:1453. DOI:10.1016/s0140-6736(07)61602-x
87. Langan S. M., Schmidt S. A., Wing K., et al. (2018). The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ* **363**:k3532. DOI:10.1136/bmj.k3532
88. Benchimol E. I., Smeeth L., Guttman A., et al. (2015). The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLoS Med.* **12**:e1001885. DOI:10.1371/journal.pmed.1001885
89. Totik N., Yücel Karakaya S. P. and Alparslan Z. N. (2023). An introduction to propensity score analysis: Checklist for clinical researches. *Eur. J. Therapeutics* **29**:667–676. DOI:10.58600/eurjther1813
90. Skrivankova V. W., Richmond R. C., Woolf B. A. R., et al. (2021). Strengthening the

- reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): Explanation and elaboration. *BMJ* **375**:n2233. DOI:10.1136/bmj.n2233
91. Walker V., Sanderson E., Levin M. G., et al. (2024). Reading and conducting instrumental variable studies: Guide, glossary, and checklist. *BMJ* **387**:e078093. DOI:10.1136/bmj-2023-078093
92. Correia L. C. L., Mascarenhas R. F., De Menezes F. S. C., et al. (2025). Confounder selection in observational studies in high-impact medical and epidemiological journals. *JAMA Netw. Open* **8**:e2524176. DOI:10.1001/jamanetworkopen.2025.24176
93. Wei Q., Cui M., Liu Z., et al. (2025). Integrating statistical design and inference: A roadmap for robust and trustworthy medical AI. *Innov. Med.* **3**:100145. DOI:10.59717/j.xinn-med.2025.100145

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (82273733), Shanghai Shengkang Hospital Development Center Projects (SHDC2023CRSO34), and 2025 Xuhui District Hospital-Local Collaborative Projects (25XHYD-04). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

L.H. and G.F. jointly conceptualized the research theme and core framework. Y.W. and W.L. refined the specific content and jointly drafted the manuscript. L.W., Y.W. and Y.Z. refined the academic details and reviewed the professional issues. All the authors critically reviewed and revised the manuscript. All authors contributed to and approved the manuscript.

DECLARATION OF INTERESTS

Guoshuang Feng is an Editorial Board member of The Innovation Medicine and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and their research group. The other authors declare no conflicts of interest.

DATA AND CODE AVAILABILITY

Not applicable.