Industry Update

For reprint orders, please contact: info@becaris.com

Journal of **Comparative Effectiveness Research**

R WE ready for reimbursement? A round-up of developments in real-world evidence relating to health technology assessment: part 26

Paul Arora[1,2] & Sreeram V Ramagopalan*[,3]

[1]Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, ON M5T 3M7, Canada

[2]Inka Health, 100 King St W. Suite 1600. Toronto, ON M5X 1G5, Canada

[3]Centre for Pharmaceutical Medicine Research, King's College London, London, SE1 9NH, UK

*Author for correspondence: sreeram@ramagopalan.net

In this update, we review a framework for identifying and mitigating information bias in electronic health records and administrative claims data, highlighting practical recommendations for study design, variable definition, and statistical analysis. We also discuss a perspective on emerging privacy-preserving technologies – synthetic data and federated networks – that enable secure cross-border data access while maintaining patient privacy.

First draft submitted: 21 March 2026; Accepted for publication: 8 April 2026; Published online: 5 May 2026

Keywords: claims data • data access • electronic health records • federated networks • health technology assessment • information bias • MHRA • patient privacy • real-world data • real-world evidence • synthetic data

Information bias is a threat to the validity of real-world evidence (RWE), arising from the inaccurate measurement or misclassification of treatments, outcomes or confounders [1]. This is a particularly salient concern in real-world data (RWD) sources that were collected for purposes other than research. However, guidance on measurement error and misclassification in RWD remains underdeveloped and Arena and colleagues sought to address this gap by undertaking a targeted literature review in order to develop a framework to help researchers identify and address information bias [1]. Arena *et al.* searched PubMed for articles published between January 2019 and May 2024, focusing specifically on RWD investigations that explicitly contained information bias case studies and review articles directly addressing information bias. The search yielded 294 unique records, of which 38 articles and guidance ultimately met inclusion criteria after rigorous screening. Notably, the authors also incorporated the US FDA's 2024 guidance on assessing electronic health records (EHRs) and medical claims data [2]. The identified studies spanned diverse therapeutic areas, with infectious disease (n = 8), oncology (n = 6), cardiovascular disease (n = 4) and mental health (n = 3) being most represented. The majority of studies were conducted in the USA (n = 25), though investigations from Australia, Canada, Latin America, South Africa and the UK were also included. EHR data represented the most commonly used RWD source, either alone (n = 24) or in combination with claims data (n = 7). From the 38 included articles and guidance, Arena *et al.* extracted 82 specific recommendations and distilled these into 15 general recommendations organized across three protocol design elements: study design (six recommendations), study variables (four recommendations) and statistical analyses (five recommendations). For study design, data linkage emerges as a key mitigation strategy across multiple studies. Linking patient-level data from multiple sources – such as combining claims data with vaccine registries or EHR data with claims data – provides a more comprehensive understanding of patients, their health status, and healthcare utilization, thereby minimizing misclassification [3]. The framework also emphasizes validation studies to validate algorithms and guard against variable misclassification, with the FDA guidance strongly advocating for this approach. Additional strategies include innovative sampling approaches for validation (e.g., multiwave sampling), data visualization to identify potential misclassification, assessment of observability impacts and thoughtful comparator group selection to address bias such as surveillance bias. Regarding study variables, the framework highlights the importance of

using validated algorithms – either externally or internally validated – to define study variables. The incorporation of clinical practice information when defining variables is emphasized, such as considering appropriate gap days between prescriptions when defining drug exposure or accounting for free samples that may not appear in records. The framework also recommends analytical techniques to refine variable definitions where needed, such as K-means clustering for better assessment of time-varying drug use patterns. For statistical analyses, quantitative bias analysis (QBA) to evaluate information bias impact emerges as a dominant theme [4]. Multiple studies have demonstrated the versatility of QBAs in quantifying bias extent, adjusting for its effects, and improving interpretation of findings [5–7]. The framework also highlights multiple imputation, Bayesian methods, and calibration methods for addressing missing data and misclassification. For manufacturers, this work has several implications. First, the framework provides a structured approach to proactively addressing information bias during study design rather than treating it as a *post hoc* limitation. By implementing the framework's recommendations – such as conducting validation studies, linking complementary data sources, and employing QBA – manufacturers can strengthen the credibility of their RWE submissions. Second, the framework highlights that not all mitigation strategies require equal resources; some approaches, such as using validated algorithms or incorporating clinical practice knowledge into variable definitions, may be more feasible than others like extensive data linkage projects. Manufacturers should clearly articulate their bias mitigation approaches in health technology assessment (HTA) submissions, demonstrating awareness of potential limitations and the steps taken to address them. The framework should be used in conjunction with HTA guidance and other tools such as the HARPER protocol template, PRINCIPLED and APPRAISE [8–10]. As RWE continues to play an expanding role in HTA decision-making, ultimately manufacturers who can demonstrate rigorous approaches to bias mitigation will be better positioned to have their evidence accepted by decision-makers.

Building on the foundation of robust study design to minimize information bias, another challenge to grapple with is that of accessing data while ensuring patient privacy. Wang *et al.* provide an examination of two emerging privacy-preserving technologies – synthetic data and federated networks – that promise to address barriers to healthcare data access [11]. As previously discussed in this series [12,13], synthetic health data represents one promising approach to addressing privacy concerns. These artificial data are intended to mimic the properties and relationships seen in real patient data, preserving statistical characteristics of the original data without containing information acquired from actual human subjects. Wang and colleagues provide examples of synthetic data applications across sectors, especially from government as one of the authors was from the UK's Medicines and Healthcare products Regulatory Agency (MHRA). As the UK's regulator of medical products, including Software as a Medical Device (SaMD) and AI as a Medical Device (AIaMD), the MHRA's interest in synthetic data originally emerged in 2017 within a specific regulatory context: the need for external validation of machine learning algorithms when alternative RWD sources were unavailable. Given the intended purpose, the MHRA's requirement was for high-fidelity synthetic data capable of capturing both the complex inter-relationships between various data fields and the statistical properties of real data. For the initial proof-of-concept project, the MHRA used an extract of anonymized, coded, tabular primary care data from the Clinical Practice Research Datalink (CPRD), which is the MHRA's own RWD, as the ground truth data for generating synthetic data. The team developed an evaluation framework to assess three critical dimensions of the generated synthetic data: fidelity (how well the synthetic data captured the statistical properties and relationships of the original data), utility (whether the synthetic data could support the intended analytical tasks and produce similar results to analyses on real data), and privacy (the degree to which the synthetic data protected against re-identification of individuals from the original dataset). The initial pilot successfully demonstrated that it was possible to generate clinically validated, high-fidelity synthetic patient data that could serve regulatory purposes. Clinical experts reviewing the synthetic data confirmed that it exhibited realistic patterns of diagnoses, prescriptions and healthcare utilization that would be expected in actual primary care populations. A notable methodological contribution from the MHRA's work was their preferential adoption of Bayesian network approaches to synthetic data generation over generative adversarial network-based approaches. This decision was driven primarily by explainability considerations: Bayesian networks were more interpretable and explainable to clinical experts undertaking the clinical validation of the synthetic datasets. When clinicians reviewed synthetic patient records, they could better understand how the Bayesian network had modeled the conditional dependencies between variables, making it easier to identify whether the synthetic data generation had properly captured genuine clinical relationships or had introduced spurious associations. The MHRA further refined the synthetic data generation approach to address two specific challenges prevalent in health data. First, they developed methods to handle the temporal nature of health datasets, recognizing that patient records contain longitudinal

information where the sequence and timing of events matter for clinical interpretation. Second, they addressed the issue of missing fields in the ground truth data, a ubiquitous challenge in real-world health data where incomplete recording is common.

Beyond improving the technical quality of synthetic data generation, the MHRA has contributed to advancing methodology for addressing representation bias. They developed a novel approach to detecting biases due to underrepresentation in real data using uncertainty analysis, then correcting these biases via conditional boosting of underrepresented groups using synthetic data. More recently, the MHRA has been researching the application of high-fidelity synthetic data in the context of clinical trials, including two particularly innovative use cases. First, they are exploring data augmentation to boost small sample sizes, which could be especially valuable in rare disease trials where recruitment challenges often result in studies that lack statistical power to detect meaningful treatment effects. Second, they are investigating synthetic control arms as an alternative to traditional external controls. A validation study of these clinical trial applications is currently underway.

Federated data networks represent a complementary privacy-preserving paradigm that allows data to remain under the control of original stewards behind their firewalls while enabling users to run analyses across multiple sites without centralizing data. Notable examples include the FDA-funded Sentinel System, the Patient-Centered Outcomes Research Institute's PCORnet, Canada's CNODES, the US-based OHDSI and the EMA's DARWIN-EU network. The analytical sophistication possible within federated data networks has evolved considerably from simple descriptive queries to inferential analyses. For descriptive studies that aim to examine utilization patterns of medical products or the natural history of diseases, analyses can generally be performed using only summary-level information exchanged between sites. This approach has been extensively used in Sentinel to understand disease epidemiology and medication utilization patterns, including studies investigating the use of systemic corticosteroids for COVID-19 in outpatient settings. For inferential studies that aim to assess causal effects of medical treatments – the types of comparative effectiveness and safety studies most relevant for HTA decision-making – methodological innovations have enabled increasingly sophisticated analyses within federated architectures. Historically, meta-analysis of database-specific results was the primary analytic option: each participating site would independently conduct a complete analysis (including confounding adjustment, outcome modeling and effect estimation), then site-specific effect estimates would be combined using meta-analytic methods. While this approach enables causal inference without sharing individual-level data, it has limitations. Meta-analysis of site-specific estimates can be inefficient when individual sites have small sample sizes, and it may not fully account for heterogeneity across sites in ways that more integrated analyses could address. More recently, several federated data networks have exploited the mathematical properties of summary scores – particularly propensity scores and disease risk scores – to perform analyses without sharing individual-level data. The key insight is that propensity scores (the probability of treatment given covariates) and risk scores (the probability of the outcome given covariates) are sufficient statistics for confounding adjustment under certain conditions. This means that if sites can share distributions of these scores rather than individual patient data, the coordinating center can perform matching, stratification, or weighting across the entire federated population. This approach has been implemented in Sentinel to examine comparative safety of medical products, including the risk of venous thromboembolism associated with oral contraceptives. These applications demonstrate that federated networks can support the types of rigorously adjusted comparative effectiveness analyses that payers require for decision-making. However, certain complex analyses may still require transfer of individual-level data. For instance, analyses requiring sophisticated adjustment for time-varying covariates where treatment and confounders change over the course of follow-up. Another limitation is that federated networks may also implement additional privacy safeguards – for example, cell suppression rules prevent sharing of small counts that could potentially identify individuals, typically requiring minimum cell sizes of five or ten patients before results can be released.

Synthetic data provides a pathway to overcome data access barriers that have historically constrained RWE studies, particularly where privacy regulations are stringent or data access negotiations are prohibitive. However, manufacturers must recognize that synthetic data requires substantial methodological rigor. Validation across multiple dimensions – fidelity, utility and privacy – is essential. The MHRA's preferential adoption of Bayesian networks over generative adversarial networks for their explainability to clinical validators illustrates an important principle: generation methods must be transparent and defensible to nontechnical stakeholders. As previously noted in this series, simply asserting that synthetic data 'looks similar' to real data will be insufficient; quantitative evaluations of how well synthetic data reproduces key statistical properties and whether analyses on synthetic data yield comparable results to real data analyses will be essential. Federated networks offer different advantages: access

10.57264/cer-2026-0074

to tens or hundreds of millions of patients across multiple healthcare systems and countries without negotiating individual data agreements. This scale and diversity is invaluable for rare diseases/outcomes and analytical methods such as propensity score approaches and distributed regression now enable adjusted comparative effectiveness analyses without centralizing patient data. Critically, privacy preserving technologies do not substitute for rigorous study design. The comprehensive framework developed by Arena *et al.* for addressing information bias – emphasizing validation studies, data linkage, thoughtful variable definition, and QBA – remains essential regardless of data access approach. Manufacturers who successfully integrate privacy-preserving technologies with robust bias mitigation methodologies, maintain transparency with decision makers about methods and limitations, and contribute to validation research will be best positioned to generate credible, diverse RWE supporting patient access to innovative therapies.

Financial disclosure

Author SV Ramagopalan has received an honorarium from Becaris Publishing for the contribution of this work. The authors have received no other financial and/or material support for this research or the creation of this work apart from that disclosed.

Competing interests disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Writing disclosure

No writing assistance was utilized in the production of this manuscript.

References

1. Arena PJ, Sun Y, Jaksa A *et al.* Information bias in electronic health records and administrative claims data: a targeted review of the recent literature. *Clin. Pharmacol. Ther.* 119(2), 362–367 (2026).
2. Castanon A, Tsvetanova A, Ramagopalan SV. RWE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 16. *J. Comp. Eff. Res.* 13(8), e240095 (2024).
3. McDonald L, Schultze A, Carroll R, Ramagopalan SV. Performing studies using the UK Clinical Practice Research Datalink: to link or not to link? *Eur. J. Epidemiol.* 33(6), 601–605 (2018).
4. Leahy TP, Kent S, Sammon C *et al.* Unmeasured confounding in nonrandomized studies: quantitative bias analysis in health technology assessment. *J. Comp. Eff. Res.* 11(12), 851–859 (2022).
5. Gupta A, Hsu G, Kent S *et al.* Quantitative bias analysis for single-arm trials with external control arms. *JAMA Netw. Open* 8(3), e252152 (2025).
6. Wilkinson S, Gupta A, Scheuer N *et al.* Assessment of alectinib vs ceritinib in ALK-positive non-small cell lung cancer in Phase II trials and in real-world data. *JAMA Netw. Open* 4(10), e2126306 (2021).
7. Popat S, Liu SV, Scheuer N *et al.* Addressing challenges with real-world synthetic control arms to demonstrate the comparative effectiveness of pralsetinib in non-small cell lung cancer. *Nat. Commun.* 13(1), 3500 (2022).
8. Simpson A, Ramagopalan SV. R WE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment. *J. Comp. Eff. Res.* 10(10), 797–799 (2021).
9. Simpson A, Ramagopalan SV. R WE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 9. *J. Comp. Eff. Res.* 11(16), 1147–1149 (2022).
10. Castanon A, Bray BD, Ramagopalan SV. R WE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 15. *J. Comp. Eff. Res.* 13(5), e240033 (2024).
11. Wang EH, Myles P, Foraker R *et al.* Crossing borders securely: synthetic data and federated networks for privacy-preserving access to real-world data and emerging use cases. *NPJ Digit. Med.* 8(1), 758 (2025).
12. Arora P, Ramagopalan SV. R WE ready for reimbursement? A round-up of developments in real-world evidence relating to health technology assessment: part 22. *J. Comp. Eff. Res.* 14(12), e250149 (2025).
13. Arora P, Ramagopalan SV. R WE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 18. *J. Comp. Eff. Res.* 14(4), e250014 (2025).