Real-world data in clinical development: Statistical considerations and reporting challenges

Natasa Rajicic[1], **Anders Mattsson**[2], **Maria Kołtowska-Häggström**[3]
[1] SC Rules, LLC, Warwick, NY, USA
[2] Stockholm, Sweden
[3] Proper Medical Writing, Warsaw, Poland

doi: 10.56012/kygt4232

Correspondence to:
Natasa Rajicic
natasa.rajicic@outlook.com

Abstract
Real-world data have an increasingly important role in clinical development and regulatory decision-making. When incorporated correctly, they can provide a unique and valuable insight into patient populations, treatment patterns, and health outcomes in support to the traditional clinical development. To that end, transparency in reporting, including clear documentation of study populations, data sources, statistical methods, and limitations, is critical, particularly when seeking regulatory acceptance. Recognised standards of reporting should be considered as they can help to enhances reproducibility and regulatory acceptance.

Introduction
Real-world data (RWD) and real-world evidence (RWE) have long been utilised for a variety of purposes such as characterisation of population health and disease trends or to study risk associated with different exposures, just to name a few. RWD and RWE have also been an important part of drug safety surveillance, especially following a market drug approval after which a new medicine starts to be used in clinical practice. More recently, RWD and RWE are increasingly used in clinical development and regulatory decision-making in new drug applications. While the conventional clinical trials, randomised clinical trials (RCTs) in particular, have long been a cornerstone of clinical development programmes and regulatory submissions due to their rigorous designs that enable causal interpretations, real-world data is becoming an important supplementary source of evidence in clinical development of new medicines and regulatory approval decisions.

The term *RWD* refers to data derived from sources that are outside of the conventional clinical trials, including, for example, electronic health records (EHRs), medical claims databases, patient registries, and wearable health technologies.[1] Evidence generated from RWD studies has been used to inform disease histories, safety surveillance in post-marketing, quality of life outcomes, or treatment effectiveness in clinical practice, just to name a few.[2,3] Recently, the integration of RWD into clinical development has been recognised and promoted by regulatory agencies such as the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA), for their potential to support and inform drug approvals and policy decisions.[4,5]

While any research design is concerned with issues such as selection bias, confounding, and variabilities in data collection and endpoint definitions, these concerns are even more prominent when it comes to RWD. Anticipating and addressing these issues is crucial for maintaining the integrity, validity, and applicability of findings that result from RWD. Transparent reporting contributes to the credibility and validity of findings, particularly when RWD are used as a part of regulatory submissions.

Transparent reporting contributes to the credibility and validity of findings, particularly when RWD are used as a part of regulatory submissions.

This article provides an overview of some common statistical methodologies employed when analysing real-world data, and discusses the challenges associated with reporting RWD findings, with a particular emphasis on statistical and interpretation issues essential for maintaining methodological integrity, validity, and transparency when reporting analyses of RWD.

Why RWD in clinical development
RWD have long been used to support post-marketing safety surveillance, continued benefit-risk evaluations, and label extension applications, e.g., in rare diseases that have limited patient populations or in situations where traditional clinical trials would be unfeasible or unethical. An increasingly attractive use of RWD is within the clinical development phase, where for example, a traditional control group would be impractical or unethical. Here, RWD are used as a source for creating an external control group, thus enabling a structured comparator where otherwise one would be absent from the investigation.

In fact, several features of traditional clinical trials, and RCTs in particular, make the use of RWD an attractive fit complementing clinical development. For example, clinical trials often have strict inclusion and exclusion criteria as these can help create a more homogeneous study population that can in turn reduce overall variability and increase precision and power of estimation. Homogeneity of the study population can also reduce the impact of known and unknown confounding variables. However, strict patient selection criteria can make the study interpretation less generalisable to real-world clinical settings. Rare disease studies with limited patient population pools often face challenges in enrolling enough participants for an adequately designed and powered study. Rigorously designed, monitored, and executed studies are often prohibitively expensive and essentially impossible to carry out unless conducted by large pharmaceutical sponsors or consortia. Interventions that are studied in highly controlled clinical trials that do not mirror routine clinical practice limit the generalisability of the results.

Image: Freepik

Finally, clinical trials may not be conducted over a long enough period to provide data on either late-emerging adverse events or effects in incurable chronic conditions that require life-long treatment.

Given these limitations, integrating RWD alongside RCTs can enhance evidence generation by providing insights into broader patient populations, long-term effects, and real-world treatment effectiveness.[6] Therefore, appropriately designed and executed analyses based on RWD can be a complementary and useful tool in filling the gaps present in traditional clinical trials.

Commonly used statistical methodologies in RWD analysis

Analyses and inferences based on RWD often require different statistical methods compared to the analytical approaches used in analyses of typical clinical trial, especially when contrasted with RCTs. This is because most statistical methods used in standard analyses assume that the patient groups being compared are reasonably well balanced – both in terms of known and unknown potential confounders – prior to the introduction of the intervention. They also assume that follow-up of participants remains comparable across groups, except for differences directly attributable to the intervention itself. Use of randomisation can help with the first issue, and principled adherence to a well-developed protocol that strives for controlled and uniform follow-up procedures can help deal with the second issue. Consequently, results of the statistical tests evaluating the differences among study groups can be potentially interpreted as causality.

The absence of the above two and other considerations when conducting the analyses based on RWD, as a result, necessitate application of statistical method that are sensitive to such issues. The following is a summary of few statistical approaches that were primarily developed to handle data outside of RCTs and that can be found useful when analysing RWD.

Regression analysis models

Regression analysis is a key statistical tool used to explore and quantify relationships between variables, such as treatment exposure and clinical outcomes. In epidemiological studies and RWD analyses, regression models such as linear and logistic regression, repeated measures analysis, or Cox proportional hazards modes, are commonly applied to control for confounding factors when characterising relationships between exposure and responses. By adjusting for patient demographics, comorbidities, and other covariates, regression can help isolate the impact of a specific variable of interest, enhancing the validity of conclusions drawn from non-randomised, real-world settings. However, regression analysis in RWD has limitations, including confounding due to unmeasured or misclassified variables, model misspecification, selection bias, or missing data. Nevertheless, regression methods, especially when used properly, remain an essential and widely used tool for the statistical analyses in RWD settings.

By adjusting for patient demographics, comorbidities, and other covariates, regression can help isolate the impact of a specific variable of interest.

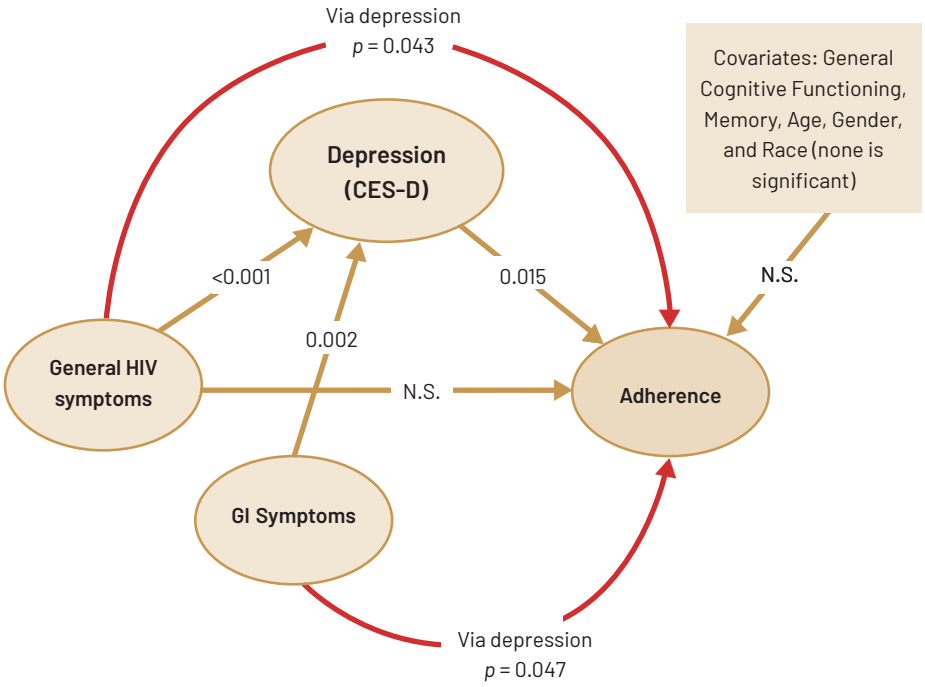

Figure 1. Graphical illustration of use of structural equation models to characterise and evaluate causal relationships involving direct and indirect effects among variables

The graph indicates that both general HIV-related symptoms and gastrointestinal (GI) symptoms are directly associated with higher levels of depressive symptoms. In turn, higher levels of depressive symptoms are directly linked to lower medication adherence. Notably, general HIV-related symptoms do not have a direct effect on adherence; rather, their impact is indirect, mediated by depressive symptoms. This suggests that an increase in general HIV-related symptoms is associated with increased depressive symptoms, which subsequently lead to poorer adherence. Similarly, GI symptoms also exert an indirect effect on adherence through depressive symptoms, with no direct relationship observed between GI symptoms and adherence.

Propensity score methods

Propensity score methods, such as propensity score matching and inverse probability weighting, were developed to help create groups of patients that are balanced with respect to observed baseline characteristics in observational studies.[7] Propensity score methods can generally be divided into two categories: propensity score matching, which attempts to pair patients with similar characteristics across different treatment arms, and inverse probability of treatment weighting, which assigns weights to patients based on their propensity scores to create a pseudo-randomised population. In this way, propensity scores construct pools of patients that appear similar with respect to the distributions of covariates, irrespective of the actual treatment subsequently received. Thus, if properly executed, analysis of the differences in outcomes between treatment groups that incorporate propensity scores could help adequately evaluate treatment differences in RWD analyses.

Causal inference methods

The topic of causal inference is rich and long-standing. Numerous methods and relevant theories have been developed.[8] Moreover, causal inference methods are rarely a standard topic in statistical academic programmes, even in advanced post-graduate studies, and like many advanced methodologies, they should be handled by experienced professionals only. Many of these approaches aim to examine relationships between variables or concepts that are not directly observable from the data or attempt to estimate causal relationships from data in the presence of various types of confounding. Here we introduce only a few methods as an illustration; more comprehensive reviews can be found elsewhere.[8,9]

Structural equation modelling (SEM) provides a framework for examining relationships between observed variables and underlying constructs – latent variables – that cannot be directly observed (e.g., depression or quality of life) but are inferred from other measurable variables. SEMs utilise and combine methods of factor analysis and regression and can be visualised using diagrams that depict hypothesised causal directional paths among variables (Figure 1). An example of SEM application is a study to examine the process by which direct and indirect effects of HIV-related symptoms are related to adherence to antiretroviral therapy as well as whether the symptom of depression acts as a mediator of this relationship.[10]

Bayesian methods

Bayesian statistical methods, in which inference is made based on data-driven updates to prior beliefs, has found numerous applications in the design and analysis of clinical trial data. Bayesian analysis approaches incorporate prior external information, for example evidence from completed clinical trials or expert opinion, into current analyses, enabling more robust inference even when data are limited or heterogeneous. In RWD applications, Bayesian models can account for missing data by predicting unknown values using the available data (e.g. through multiple imputation), adjust for confounding factors through Bayesian propensity score methods, or account for variability across different populations using hierarchical modelling, as in *pragmatic* trials.[11] Bayesian analysis can also facilitate dynamic updating of inferences as new data become available, making them particularly valuable for ongoing studies and real-time decision-making.[12]

Table 1. Reporting strategies for ensuring transparency

Reporting topic	What needs to be described or included?	Why?
Information on data sources and its quality	• Data origin that is sufficiently and clearly detailed, with specific sources named (e.g., EHRs, claims, registries, etc.) and whether data were collected for a specific purpose or extracted from a database that collected data without a prespecified purpose • Details on how data were extracted or collected, managed, cleaned, and processed, including what steps were taken to maintain data integrity and quality	• Details around the processes of data collection and curation can provide important insights into any limitations or potential biases inherent to the data • A reader should be sufficiently informed prior to making decisions and/or interpretations based on the results
Sources of patient population	• Inclusion and exclusion criteria with attention to any specific characteristics • For registries, selection criteria described separately for the entire registry and for the subset of patients analysed in a specific registry-based study • When RWD are used to supplement data from clinical trials, detailed differences in the populations • Clear distinction between e.g., mining of the entire registry vs. targeted selection of data from a registry based on pre-specified inclusion/exclusion criteria	• Understanding the source population helps reviewers understand the generalisability of the findings and how representative results are of any targeted populations • Without randomised assignment to treatment, participants who are treated may be inherently and systematically different from those who are not
Analysis plans and methods	• Comprehensive description of the statistical methods used, all assumptions clearly stated • Justification of methodological choices and any alternative strategies considered • Report of sensitivity analyses conducted to assess robustness of findings	• Transparency regarding analytical methods used ensures reproducibility, a basic tenant of rigorous research • Regulatory agencies require transparency in how RWE studies are conducted
Sources of bias and confounding	• Potential sources of bias, such as selection bias and confounding, and how adjustments were made to minimise these biases • Results presented both before and after adjustment to illustrate their impact	• The source of the RWD is in routine clinical practice where factors like disease history, prior treatments, and clinical settings can influence outcomes and introduce bias when interpreting treatment effectiveness
Consistency in endpoint definitions	• Clinical endpoint definitions standardised across different data sources, including sites and institutions • Criteria used for endpoint derivation and the validation process	• Different RWD sources may define clinical endpoints differently which can translate into a different outcome or endpoint when it comes to the analysis • The degree of consistency in endpoint definitions is essential for interpretation of the results

Machine learning methods

Finally, in this brief overview of statistical methods, machine learning should be mentioned as well since RWD often involve large, complex datasets where machine learning can be particularly useful. The results include improved insight through supervised learning (e.g., Random Forests) or assistance in predictive modelling and treatment effect estimation.[13] Similarly, unsupervised learning such as clustering, or dimensionality reduction (e.g., principal component analysis or lasso regression) can help identify patterns within patient populations.[14]

Reporting of RWD

Ensuring transparency in reporting

Transparency is vital for ensuring credibility, reproducibility, and ultimately regulatory acceptance of analyses based on RWD. Table 1 highlights several important objectives when summarising evidence arising from RWD.

In addition to the statistical reporting issues listed in Table 1, a transparent reporting of RWD should include topics of missing data. Namely, these include quantification of the extent of missing data and its potential impact on interpretation of analysis results, methods used to handle missing data in the analysis (e.g., only complete cases analysed, or type of imputation method employed), as well as any sensitivity analyses completed to explore the influence of missing data assumptions on analysis results. In addition, whenever possible, access to study protocols and analytical code should be provided as this greatly enhance reproducibility. Open-source platforms are a good place for sharing, provided they adequately safeguard patient privacy.

> Bayesian analysis approaches incorporate prior external information, for example evidence from completed clinical trials or expert opinion, into current analyses, enabling more robust inference even when data are limited or heterogeneous

Importance of objective communication of findings

Like any study or data reporting, objective reporting and interpretation of findings is essential. (See, for example, the ClinicalTrials.gov repository.)[15] Given the inherent difficulties in establishing causal inference in results based on RWD, the importance of careful reporting of RWD analyses should be emphasised. A careful consideration not to overstate causal relationships, especially given the observational nature of RWD studies, is essential. Communication of RWD findings requires measured and balanced language with ample discussion of potential biases, while acknowledging any limitations. Remaining uncertainties should be highlighted, and, when meaningful, alternative explanations to the findings presented.[16]

Enhancing transparency not only strengthens confidence in RWE but also facilitates its successful integration into clinical decision-making and regulatory assessments. This not only builds confidence among regulators and the scientific community but also paves the way for the broader integration of RWD in healthcare.

Conclusions

RWD will increasingly be used to complement traditional clinical trial data. Their applications – from constructing external control arms and enhancing safety surveillance to supporting research in rare diseases, underscore their growing role in clinical development and regulatory decision-making. However, the inherent challenges of RWD, such as selection bias, confounding, and inconsistencies in data collection, require specialised analytical approaches such as propensity score methods, Bayesian techniques, and causal inference models. Additionally, transparency in reporting, including clear documentation of study populations, data sources, statistical methods, and limitations, is critical to maintaining scientific rigor, particularly when seeking regulatory acceptance. Recognised standards of reporting, such as those outlined by STROBE and RECORD,[17,18] should be considered as they can help to enhance reproducibility and regulatory acceptance.

This article has focused on statistical and reporting considerations, but numerous other aspects of RWE should be considered, including appropriate regulatory frameworks. Fortunately, significant strides in formalising common practices and providing industry guidance have already been accomplished.[4,5] Other important considerations that need addressing are data privacy, and ethical and security concerns when utilising RWD.[6] The growing collaboration across industry, academia, and regulatory bodies is encouraging and welcomed and will likely lead to industry-wide, recognised best practices towards greater utilisation of the RWD and RWE.

Image: Freepik

Disclosures and conflicts of interest

The authors declare no conflicts of interest.

References

1. Liu F, Panagiotakos D. Real-world data: brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol. 2022;22(1):287. doi:10.1186/s12874-022-01768-6
2. Dang A. Real-world evidence: A primer. Pharmaceut Med. 2023;37(1):25–36. doi:10.1007/s40290-022-00456-6
3. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. JAMA. 2018;320(9):867–8. doi:10.1001/jama.2018.10136
4. US Food and Drug Administration. Use of real-world evidence to support regulatory decision-making for medical devices. August 31, 2017. FDA-2016-D-2153
5. EMA. Journey towards a roadmap for regulatory guidance on real-world evidence. 19 February 2025. EMA/100019/2025
6. US Food and Drug Administration. Framework for FDA's Real-World Evidence Program. 2018 [cited 2025 Aug 25]. Available at: https://www.fda.gov/media/120060/download
7. McMurry TL, Hu Y, Blackstone EH, et al. Propensity scores: methods, considerations, and applications. J Thorac Cardiovasc Surg. 2015;150:14–19. doi:10.1016/j.jtcvs.2015.03.057

8. Brumback BA. Fundamentals of Causal Inference: With R (1st ed.). Chapman and Hall/CRC. 2021. doi:10.1201/9781003146674
9. Dang LE, Gruber S, Lee H, et al. A causal roadmap for generating high-quality real-world evidence. J Clin Transl Sci. 2023;7(1):e212, 1–12. doi:10.1017/cts.2023.635
10. Yoo-Jeong M, Waldrop-Valverde D, McCoy K, et al. A structural equation model of HIV-related symptoms, depressive symptoms, and medication adherence. HIV AIDS. 2016:2(3). doi:10.16966/2380-5536.123
11. Le-Rademacher J, Gunn H, Yao X, Schaid DJ. Clinical trials overview: from explanatory to pragmatic clinical trials. Mayo Clin Proc. 2023;98(8):1241–53. doi:10.1016/j.mayocp.2023.04.013
12. Edwards D, Best N, Crawford J, et al. Using Bayesian dynamic borrowing to maximize the use of existing data: a case-study. Ther Innov Regul Sci. 2024;58(1):1–10. doi:10.1007/s43441-023-00585-3
13. Emir B, Johnson K, Kuhn M, Parsons B. Predictive modeling of response to pregabalin for the treatment of neuropathic pain using 6-week observational data: a spectrum of modern analytics applications. Clin Ther. 2017;39(1):98–106. doi:10.1016/j.clinthera.2016.11.015
14. Wang J, Xu Y, Liu L., et al. Comparison of LASSO and random forest models for predicting the risk of premature coronary artery disease. BMC Med Inform Decis Mak. 23(1):297. doi:10.1186/s12911-023-02407-w
15. National Institutes of Health (US). Bethesda (MD): National Library of Medicine (US). Available from: https://clinicaltrials.gov
16. Verweij S, Weemers J, Jonker C, et al. Learning from European regulator-initiated studies for regulatory decision making. Drug Discov Today. 2025;30(1):104256. doi:10.1016/j.drudis.2024.104256
18. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344–9. doi:10.1016/j.jclinepi.2007.11.008
19. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015 Oct 6;12(10):e1001885. doi:10.1371/journal.pmed.1001885

Author information

Natasa Rajicic, ScD, has been a practicing biostatistician for more than 20 years. Her expertise includes adaptive study designs and implementation, oncology, and rare disease. She has served as an independent statistical reviewer as a member of monitoring boards. Earlier in her career, Natasa was a biostatistician at Arrowhead Pharmaceuticals, Cytel, Pfizer, Center for Biostatistics in AIDS Research in Boston, and taught Research Design at Columbia University in New York.

Anders Mattsson, PhD, has experience for more than 30 years in the field of biostatistics. Special focus has been on epidemiology and Phase III and Phase IV clinical trials. Special areas of interest have been on study design, statistical methods in epidemiology, ionising radiation epidemiology, pharmacoepidemiology, risk estimation and effectiveness analyses in observational studies. He has authored or co-authored over 70 peer-reviewed publications. He worked at the Karolinska Institute, for the regional cancer registry in Stockholm, and in the pharmaceutical industry before he retired in 2020.

Maria Kołtowska-Häggström, MD, PhD, runs Proper Medical Writing, the first Polish medical writing agency. She has 30 years of experience in the pharmaceutical industry, primarily in clinical research, including real-world data (RWD) and real-world evidence (RWE). Maria has authored over 80 peer-reviewed publications, many focused on RWE and patient-reported outcomes. She is a member of EMWA, EASE, ESE, and GRS; serves as a journal reviewer and Associate Editor for BMC Endocrine Disorders; and is an EMWA workshop leader and a section editor for *Medical Writing*.