A year in review:

real-world data, decision-making and COVID-19

Five FREE journal articles

Content includes:

Special Report: Replication of randomized clinical trial results using real-world data: paving the way for effectiveness decisions

Commentary: Real-world evidence and nonrandomized data in health technology assessment: using existing methods to address unmeasured confounding?

Commentary: Life after COVID-19: R WE going to help?

White Paper: Strengthening pharma's contract with society: the value of trusted partnerships between pharma and healthcare facilitated by real-world data

Review: Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants

The Evidence Base®

Special Report

Journal of Comparative Effectiveness Research

Replication of randomized clinical trial results using real-world data: paving the way for effectiveness decisions

Kristin M Sheffield*[,1], Nancy A Dreyer[2], James F Murray[1], Douglas E Faries[3] & Megan N Klopchin[1]

[1]Global Patient Outcomes & Real World Evidence, Eli Lilly & Company, Indianapolis, IN 46225, USA
[2]IQVIA Real-World & Analytic Solutions, Cambridge, MA 02139, USA
[3]Real World Analytics, Eli Lilly & Company, Indianapolis, IN 46225, USA
*Author for correspondence: Tel.: +1 317 503 8821; Sheffield_kristin_m@lilly.com

The FDA is preparing guidance about using real-world evidence (RWE) to support decisions about product effectiveness. Several ongoing efforts aim to replicate randomized clinical trial (RCT) results using RWE with the intent of identifying circumstances and methods that provide valid evidence of drug effects. Lack of agreement may not be due to faulty methods but rather to the challenges with emulating RCTs, differences in healthcare settings and patient populations, differences in effect measures and data analysis, bias, and/or the efficacy–effectiveness gap. In fact, for some decisions, RWE may lead to better understanding of how treatments work in usual care settings than a more constrained view from RCTs. Efforts to reconcile the role and opportunities for generating complementary evidence from RWE and RCTs will advance regulatory science.

First draft submitted: 30 July 2020; Accepted for publication: 26 August 2020; Published online: 11 September 2020

Keywords: causal inference • real-world evidence • regulatory decision-making • replication

Background

Regulators and other decision makers have primarily relied on evidence from randomized clinical trials (RCTs) to support drug effectiveness determinations. RCTs provide a straightforward and well-known approach to minimize bias between treatment groups, as well as tightly controlled measurements to maximize data quality. These features provide *a priori* confidence that the results from well-designed and executed RCTs will be causally interpretable [1]. In contrast, evidence generated from observational studies of real-world data (RWD) is often considered inferior because nonrandom treatment assignment and less rigorous data collection may compromise internal validity, making causal interpretation of results challenging [1]. Moreover, the results of observational studies evaluating a treatment effect do not always agree with RCTs and these inconsistencies engender uncertainty and skepticism about observational research. While we may always need RCT results, a possibility exists that for some decisions, real-world evidence (RWE) can provide substantial evidence of treatment effect and may lead to a better understanding of how a treatment works in usual care settings versus a more constrained view from RCTs.

The use of RWE to inform drug effectiveness decisions has been limited to the support of approvals in rare diseases and oncology and the comparative effectiveness of preventative vaccines [2]. The FDA RWE Program plans to evaluate observational study designs and issue guidance regarding whether and how they may generate evidence to support decisions related to product effectiveness [2]. Evidence used to support product effectiveness determinations needs to establish a causal relationship between product use and healthcare outcomes [2]. While there is a considerable statistical literature on causal frameworks [3–6], standard and emerging methods for causal inference [7–9] and methods for assessing assumptions for causal inference [10], significant challenges remain. The Framework for US FDA's RWE Program suggests that replications of RCTs using RWD may provide insights into the opportunities and limitations of observational studies. Assessing the comparability of results between RCTs and rigorously designed observational studies may shed light on the clinical scenarios, study designs, end points and

Future Medicine

statistical methods that lead to comparable results, increasing confidence that these types of studies can provide reliable evidence of drug effect. In this paper we explore the issues of RCTs versus RWE replication, including clinical and statistical reasons for potential differences in results and recommendations for future efforts.

Current efforts to reconcile the role of RWE as a complement to RCT

There are several ongoing efforts that aim to replicate the results of RCTs using rigorously designed and analyzed observational studies. The results of these efforts are likely to receive considerable attention from both advocates and critics of RWE and to raise additional questions. The most prominent effort is RCT DUPLICATE, a collaboration between FDA, Brigham and Women's Hospital and Harvard Medical School Division of Pharmacoepidemiology, to replicate 30 completed Phase III or IV trials and to predict the results of seven ongoing Phase IV trials using Medicare and commercial claims data [11]. The Multi-Regional Clinical Trials Center and OptumLabs are leading another effort called Observational Patient Evidence for Regulatory Approval and Understanding Disease (OPERAND) [12]. They have funded Brown University and Harvard Pilgrim Health Care Institute to replicate the ROCKET-AF trial for atrial fibrillation and the LEAD-2 trial for Type 2 diabetes control using claims from commercial and Medicare Advantage plans and electronic medical record data from OptumLabs Data Warehouse. Finally, FDA has funded the Yale University-Mayo Clinic Center of Excellence in Regulatory Science and Innovation to predict the results of three to four ongoing safety trials using OptumLabs claims data [13].

These replication efforts will evaluate comparability of RCT and observational study results based on prespecified measures of agreement – regulatory agreement and estimate agreement [14]. Regulatory agreement indicates that the observational study replicates the direction and statistical significance of the RCT and the conclusions drawn (including decisions made) would be same [14]. Estimate agreement means that the effect estimate from the observational study lies within the 95% CI for the treatment effect from the RCT [14]. The selection of the metrics used to assess agreement matters when judging the success of these replication exercises and determining implications regarding confidence in future observational studies. Regulatory agreement seems most relevant when considering the application of observational study results to support future regulatory decisions. RCT DUPLICATE investigators expect the probability of regulatory agreement in the absence of bias to be in the 80–90% range for trials showing significant effects. For trials that failed to find significant effects, the expected probability of regulatory agreement is 95% for truly negative trials and <20% for false negative trials [14]. However, investigators have not prespecified an *a priori* expectation of regulatory agreement for the overall set of 30 studies. Such a quantification of the expected value of agreement statistic may be helpful when evaluating the results of these exercises and could be set by considering how well RCTs replicate each other according to these agreement measures.

The purpose of these replication exercises is to identify the clinical scenarios (e.g., indications and outcomes), study designs and analytic approaches that lend themselves to valid study implementation with RWD. The premise is that developing such an empirical evidence base will enable regulators to appraise with greater confidence when observational studies may be used to support regulatory decision making [14,15]. When considering these replication efforts, several questions arise for relevant stakeholders and the scientific community. For example, why do some observational studies fail to replicate the effect from the RCT while others succeed? [14] When will there be sufficient empirical evidence to predict with high certainty the validity of an observational study of treatment effectiveness?

Interpreting replication results & recommendations

Even with rigorously designed observational studies intended to mimic the target RCTs, some variation in results should be expected. There are valid reasons for why RCTs and observational studies may not agree and it is important to consider possible reasons in the context of these replication exercises. While agreement between RCTs and observational studies may be interpreted as strengthening support for the causal relationship, the absence of concurrence does not mean one approach or the other is wrong. Previous comparisons of observational studies and RCTs have primarily attributed discrepancies in results to bias and confounding in the treatment effect estimates from the observational studies [16]. However, other factors also may be responsible, such as challenges with emulating the target RCT, differences in healthcare settings, inclusion of more vulnerable or diverse patients, differences in effect measures and data analysis and the efficacy-effectiveness gap. It is worth noting that even highly cited RCTs have been later contradicted, [17] and the reasons for failure to replicate can shed insight on the design and validity of RCTs themselves.

Differences in patient populations, end point measurement & calendar time

Despite efforts to approximate and apply the eligibility criteria from the RCT, important differences may remain between patient populations in the target RCT and the observational study. First, it can be challenging to apply trial eligibility criteria to real-world data sources and criteria that involve clinical data or physiological measurements will not be measurable in administrative claims data, or even in electronic medical records [18]. Even after applying similar eligibility criteria, there may be differences between the RCT and real-world populations in the underlying distributions of baseline patient characteristics, such as age or prevalence of comorbid conditions. RCT DUPLICATE researchers made the decision not to reweight the real-world populations to match the reference RCT population in primary analyses; however, this approach may be explored in sensitivity analyses. It is possible that the patient populations from which the real-world study samples are drawn (e.g., Medicare and commercial claims databases) may have different baseline risks for the primary outcome compared with the RCT patients or even to each other, as observed in a recent database study predicting the results of the CAROLINA trial [19]. The distributions of treatment effect modifiers also might differ between the RCT and real-world population, or there may be heterogeneity of treatment effect based on population variables that may be unknown or unaccounted for in each dataset. Finally, geographic differences need to be considered. Most pivotal, registrational RCTs are global studies and patients from the USA represent a subset of the overall study population. Replication efforts based on US data only should conduct sensitivity analyses to compare results to the treatment effect estimates for US trial patient subgroups where possible.

Differences in end point measurement also may impact agreement. The outcome definitions and severity of events that are captured in RCTs may differ from those used in observational studies based on secondary data, meaning that the end points could represent somewhat different concepts. Claims-based studies require rules or algorithms to identify cohorts, exposures and outcomes events. This speaks to the desirability of using validated algorithms with known operating characteristics in RWE studies. The capture of outcome events may only occur for events that required medical care or were queried about during a medical encounter. The surveillance periods, duration of follow-up and timing of end point measurement may differ as well. Trials evaluate end points at protocol-specified time intervals, while in clinical practice patients are monitored according to usual care, which varies across practices, health systems and regions and it is not possible to determine the intent of observed tests and procedures (e.g., routine surveillance vs suspicion of disease).

Finally, there may be calendar time differences. In some cases, the observational study will require a later time period than the corresponding pivotal Phase III RCT because the drug was either not yet on the market or only used off-label at the time the RCT was conducted. Improvements in clinical practice and changes in the treatment landscape or patient population may occur over time, which can impact treatment effect estimates.

Differences in data analysis & effect measures

When comparing treatment effect estimates from the RCTs and observational studies, it is important to consider differences in data analysis and the causal effect measures under study. RCTs typically conduct intent-to-treat analysis based on how patients were randomly assigned into treatment groups, whereas observational studies group patients into treatment groups based on what was observed (as used) rather than what was intended (intent to treat) and conduct the observational analog of a 'per protocol' analysis [20]. Direct comparisons of effect estimates between RCTs and observational studies may require reanalysis of the RCT data to estimate the per-protocol effect and make adjustments for adherence and loss to follow-up [20]. Patterns of intercurrent events such as treatment adherence, dose changes, and treatment augmentation or switching are expected to vary between RCTs and RWD. In an RCT, the study protocol defines rules for treatment changes and adherence is closely monitored, while in clinical practice there is less oversight of patients and much greater variability in treatment changes and the influence of other factors, such as cost or insurance coverage and social determinants of health. RCT DUPLICATE investigators plan to censor patients when they discontinue the study treatment or switch treatments [14]. It is not clear what other approaches may be used to address ITT versus per-protocol analysis or account for treatment changes during the follow-up period in the observational studies, but this issue highlights the potential complementarity of RCTs and RWD.

Bias & confounding

It is well known that observational studies may be affected by bias and unmeasured confounding. Real-world data sources are often missing key sociodemographic and clinical variables that may confound the treatment-outcome

relationship. There may be channeling bias related to patient, physician and healthcare system factors influencing prescribing and use of a newly approved drug [21] or other biases known to harm the internal validity of observational studies, such as confounding by indication, confounding by frailty and the healthy user effect [22]. There also may be misclassification or measurement error in the exposure and baseline predictors or covariates that can impact the estimation of treatment effects and alter their interpretation [23].

None of these issues have a simple or guaranteed solution that can offer the reassurance provided by randomization. Comparative analyses based on observational data are challenging. Standard bias control methods rely on the assumption of no unmeasured confounding, which may be untenable in many observational studies of treatment effectiveness. However, one is not without some recourse. The literature is rich with methods and evaluations of methods for adjusting for confounders, such as the use of matching, stratification, weighting and regression-based methods often through use of the propensity score [4,7,8,24–27]. Emerging methods such as doubly robust approaches, double matching (propensity and prognostic scores) and even machine learning approaches such as model averaging may provide improved estimators of treatment effects [28–31]. Recent review papers are emphasizing the use of an expanding toolkit of sensitivity analysis methods to assess the potential impact of unmeasured confounding [32–35]. It is expected that competent researchers will avoid common study design flaws [1] and apply accepted methods to adjust for measured confounding, as well as more novel methods to address unmeasured confounding. It is critical to conduct extensive sensitivity analyses where researchers diligently seek and include all factors that can address any underlying bias or confounding.

The impact of known but unmeasured confounding and measurement error could be minimized even further by the selection of real-world data sources that are 'fit for purpose' for replicating the treatment effect from the RCT. The decision to use claims data for the RCT DUPLICATE project was logistical rather than scientific. It is likely that for some RCTs, administrative claims data alone will be insufficient to support replication since it may not be able to match the inclusion or exclusion criteria or have the necessary variables to adequately find and correct the bias or confounding. Linkage with laboratory data or electronic health records data may be necessary to provide richer information. If linking is not possible, then replication in different types of real-world data sources may be helpful.

Efficacy–effectiveness gap

The efficacy–effectiveness gap refers to the longstanding observation that products perform differently in clinical practice than in RCTs. There is considerable literature on this topic [36], but it has not been directly addressed in recent discussions on replications of RCTs using RWD. Imperfections of the healthcare system and delivery of care may contribute to differences in outcomes between RCTs and observational studies [36]. In routine practice, there are barriers to accessing healthcare resources, variability in testing, diagnosis, and treatment, and physician behavior and patient adherence are not optimal [36]. True differences in outcomes may be expected between highly protocolized care of RCTs and usual care in real-world settings. The efficacy–effectiveness gap may also be the result of complex interactions between the drug's biologic effects and patient, provider and healthcare-related factors [36]. While the methodological considerations discussed earlier in this paper may partially explain the efficacy–effectiveness gap, it is also possible that observational study results represent part of a continuum of truth about a treatment effect.

Recommendations

Disagreement in results of these RCT replication efforts could be evidence of challenges with emulating the target trials using RWD, differences in the data analysis and effect measures, bias or confounding in the observational study, or the efficacy–effectiveness gap. It is likely that all these factors will contribute in some way to any observed discrepancies in effect estimates. Arguably, the most important work will involve disentangling the reasons for differences to understand the results of the exercises and the implications for confidence in the internal validity of observational studies. A recent paper by the lead investigators of RCT DUPLICATE considers the potential challenges in emulating a target trial with RWD and provides a list of measures of emulation differences related to study populations, treatment strategies and outcomes (e.g., proportion of patients with labs available and the length of follow-up in RWE vs RCT) that are observable [37]. These measures, along with targeted sensitivity analyses, may help investigators understand the potential impact of emulation differences between the RCT and the RWE replication. Lodi and colleagues [20] describe a systematic approach to improve the comparison of effect estimates from RCTs and observational studies based on RWD, based on harmonization of study protocols and

data analysis to target the same causal effect and the same estimand and sensitivity analyses to investigate remaining discrepancies.

We encourage other researchers to build upon the foundation of current RCT replication efforts and conduct and publish similar replication exercises in additional therapeutic classes, with careful consideration of data and methods in the context of fitness for purpose, with attention to describing agreement and when results are discrepant, to identifying potential explanations for differences. We recommend that future efforts are expanded to include not only health insurance claims but also electronic health records data, registries, linked data sources and other clinically rich data sources. It would be most useful to focus on clinical and regulatory contexts where observational studies could demonstrate substantial evidence of treatment effect to support product effectiveness determinations and labeling changes, especially since this is an important goal identified in the 21st Century Cures Act. While FDA has accepted observational studies to support effectiveness determinations in limited instances in the past, there are a growing number of therapeutic areas where there appears to be an acceptable level of risk, particularly in oncology, in which observational studies may provide primary or supportive evidence sufficient for a label change. These may include long-term effectiveness within an already approved indication, additional claims or end points within an approved indication, changes in the combination therapy, or changes in the indicated patient population. Finally, it will be important to consider other approaches, in addition to RCT replication efforts, to address FDA concerns regarding establishing causality with observational studies, for example, the application of methods to address unmeasured confounding and work to advance methodological and statistical approaches to support causal conclusions. There is a growing literature on causal frameworks and causal estimation in observational studies; however, there is more work to be done to develop and evaluate novel statistical approaches and to advance regulatory understanding and confidence in observational studies of treatment effectiveness. Just as the requirements for an acceptable RCT have become recognized and codified, it will be necessary to develop consensus on the characteristics of high-quality observational research that could meet the standard of an 'adequate and well-controlled investigation' in order for these studies to rise to the level of substantial evidence [38,39].

Conclusion

RCTs and RWE are complementary and each contribute valuable information about patient outcomes. The gains from the use of observational studies to support regulatory decisions could be considerable. Efforts to replicate RCTs can support the credibility of observational studies for estimating treatment effectiveness by demonstrating that they can support the same regulatory decision or causal conclusion as RCTs when the clinical setting is carefully chosen, appropriate data are selected and best practices for design and analysis are followed. However, careful review and interpretation of the results will be critical, particularly if there are discrepancies between the target RCT and the observational study. The conclusion should not be that the observational study was flawed or confounded [40]. The learning achieved through investigating the reasons for differences may improve our understanding of when trustworthy causal inferences can be made from observational data but may also provide insights on how better to design RCTs to improve their generalizability and usefulness in decision-making. Efforts to reconcile the role and opportunities for generating complementary evidence from RWE and RCTs will not only advance regulatory science but also progress the learning healthcare system. The replication of RCTs is not an end goal but rather an intermediate step on the way to making statements or hypotheses about the efficacy–effectiveness gap, optimal use of medical products in real-world settings, heterogeneity of treatment effect in various subpopulations, and long-term outcomes. These are complementary insights RWE is uniquely positioned to provide.

Future perspective

FDA guidance regarding whether and how observational study designs may generate evidence to support decisions related to product effectiveness is expected by the end of 2021. Rigorously designed and conducted observational studies may offer valuable information that complements evidence from clinical trials. RCT replication efforts, advances in statistical approaches to causal estimation of treatments effects and experience with using RWE may increase regulator understanding and confidence in observational studies of treatment effectiveness. In 5–10 years from now, we expect that observational study designs will be used often to provide evidence of product effectiveness to support label changes in appropriate regulatory and clinical circumstances.

Executive summary

Background

- The use of observational studies to inform drug effectiveness decisions has been limited in the past, primarily due challenges with causal inference.
- Replications of randomized clinical trials (RCTs) using real-world data may provide insights into the opportunities and limitations of observational studies for regulatory decision making.

Current efforts to reconcile the role of real-world evidence as a complement to RCT

- Several ongoing efforts aim to replicate the results of RCTs using rigorously designed and analyzed observational studies.
- Comparability will be evaluated based on prespecified measures of agreement: regulatory agreement and estimate agreement.
- The purpose of these efforts is to identify clinical scenarios, data sources, study designs and analytic approaches that provide reliable evidence of drug effect.

Interpreting replication results & recommendations

- Discrepancies in results between RCTs and observational studies should be expected and it is valuable to carefully consider possible reasons.
- Factors that may contribute to disagreement in results include: challenges with emulating the target RCT, differences in healthcare settings, differences in study populations and end point measurement, differences in effect measures and statistical analysis, the efficacy–effectiveness gap, and bias or confounding.
- The most challenging and informative task of these replication efforts will be to disentangle these potential reasons in order to understand the implications for confidence in observational studies.
- Future work should include clinically rich data sources and focus on clinical and regulatory contexts where observational studies could support labeling changes.

Conclusion

- RCTs and observational studies provide valuable, complementary information about patient outcomes. The gains from the use of observational studies to support regulatory decisions could be considerable.
- Careful review of the results of these replication efforts will be critical. The learning achieved may improve understanding of when trustworthy causal inferences can be made from observational studies.

Financial & competing interests disclosure

KM Sheffield, JF Murray, DE Faries and MN Klopchin are employees of Eli Lilly and Company and own stock in Eli Lilly and Company. NA Dreyer is an employee of IQVIA and accepts no personal consulting or speaking fees. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

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

References

Papers of special note have been highlighted as: • of interest

1. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clin. Pharmacol. Ther.* 102(6), 924–933 (2017).
2. U.S. Food and Drug Administration. Framework for FDA's real-world evidence program (2018). www.fda.gov/media/120060/download

• **Outlines the framework for US FDA's real-world evidence program, including commentary on observational studies and expected guidance documents.**

3. Holland PW. Statistics and causal inference. *J. Am. Stat. Assoc.* 81(396), 945–960 (1986).
4. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social and Biomedical Sciences*. Cambridge University Press, NY, USA (2015).
5. Halpern JY, Pearl J. Causes and explanations: a structural-model approach – Part I: causes. *Brit. J. Phil. Sci.* 56, 843–887 (2005).
6. Halpern JY, Pearl J. Causes and explanations: a structural-model approach – Part II: explanations. *Brit. J. Phil. Sci.* 56, 889–911 (2005).
7. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55 (1983).
8. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* 25(1), 1–21 (2010).

9. Zagar A, Kadziola Z, Lipkovch I, Faries D, Madigan D. Evaluating bias control strategies in observational studies using frequentist model averaging. *J. Biopharm. Stat.* (under review).
10. Zhang X, Faries DE, Li H, Stamey JD, Imbens GW. Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiol. Drug Saf.* 27(4), 373–382 (2018).

- **Provides a review of methods for addressing unmeasured confounding and includes a flowchart with best practice recommendations.**

11. RCT Duplicate. RCT duplicate: randomized controlled trials duplicated using prospective longitudinal insurance claims: applying techniques of epidemiology (2020). www.rctduplicate.org
12. OptumLabs. Using RWD in regulatory decision-making (2020). www.optumlabs.com/work/data-regulatory-decision.html
13. Yale University Center for Outcomes Research & Evaluation. Understanding the use of existing real-world data for medical product evaluation (2020). https://medicine.yale.edu/core/current_projects/cersi/research/
14. Franklin JM, Pawar A, Martin D *et al.* Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. *Clin. Pharmacol. Ther.* 107(4), 817–826 (2020).

- **Describes randomized clinical trial (RCT) DUPLICATE initiative, including the study design implementation process.**

15. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision-making. *Clin. Pharmacol. Ther.* 105(4), 867–877 (2019).
16. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ* 352, i493 (2016).
17. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 294(2), 218–228 (2005).
18. Bartlett VL, Dhruva SS, Shah ND, Ryan P, Ross JS. Feasibility of using real-world data to replicate clinical trial evidence. *JAMA Netw. Open* 2(10), e1912869 (2019).
19. Patorno E, Schneeweiss S, Gopalakrishnan C, Martin D, Franklin JM. Using real-world data to predict findings of an ongoing phase IV cardiovascular outcome trial – cardiovascular safety of linagliptin vs. glimepiride. *Diabetes Care* 42(12), 2204–2210 (2019).

- **Uses administrative claims data to predict results from an ongoing phase IV cardiovascular outcome trial in patients with Type 2 diabetes.**

20. Lodi S, Phillips A, Lundgren J *et al.* Effect estimates in randomized trials and observational studies: comparing apples with apples. *Am. J. Epidemiol.* 188(8), 1569–1577 (2019).

- **Describe a systematic approach to improve the comparison of effect estimates from RCTs and observational studies based on real-world data.**

21. Schneeweiss S, Gagne JJ, Glynn RJ, Ruhl M, Rassen JA. Assessing the comparative effectiveness of newly marketed medications: methodological challenges and implications for drug development. *Clin. Pharmacol. Ther.* 90(6), 777–790 (2011).
22. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med. Care* 48(Suppl. 6), S114–S120 (2010).
23. Rodriguez de Gil P, Bellara AP, Lanehart RE, Lee RS, Kim ES, Kromrey JD. How do propensity score methods measure up in the presence of measurement error? A Monte Carlo study. *Multivariate Behav. Res.* 50, 520–532 (2015).
24. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* 34, 3661–3679 (2015).
25. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat. Med.* 33, 1057–1069 (2014).
26. Bodory H, Camponovo L, Huber M, Lechner M. The finite sample performance of inference methods for propensity score matching and weighted estimators. *J. Business Econ. Stat.* 38(1), 183–200 (2018).
27. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J. Am. Stat. Assoc.* 113(521), 390–400 (2018).
28. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.* 23, 2937–2960 (2004).
29. Faries D, Zhang X, Kadziola Z, *et al. Real World Health Care Data Analysis: Causal Methods and Implementation Using SAS*. SAS Institute Inc, NC, USA (2020).
30. Leacy FP, Stuart EA. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Stat. Med.* 33(20), 3488–3508 (2020).
31. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat. Sci.* 22(4), 523–539 (2007).
32. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol. Drug Saf.* 15(5), 291–303 (2006).
33. Uddin MJ, Groenwold RHH, Ali MS *et al.* Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *Int. J. Clin. Pharm.* 38(4), 1–10 (2016).

34. Vanderweele TJ, Ding P. Sensitivity analysis in observational studies: introducing the e-value. *Ann. Intern. Med.* 167(4), 268–274 (2017).

35. Zhang X, Faries DE, Li H, Stamey JD, Imbens GW. Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiol. Drug Saf.* 27, 373–382 (2018).

36. Nordon C, Karcher H, Groenwold RHH *et al.* The "efficacy-effectiveness gap": historical background and current conceptualization. *Value Health* 19(1), 75–81 (2016).

37. Franklin JM, Glynn RJ, Suissa S, Schneeweiss S. Emulation differences vs. biases when calibrating real-world evidence findings against randomized controlled trials. *Clin. Pharmacol. Ther.* 107(4), 735–737 (2020).

- **Provides example measures that may be used to assess emulation differences between RCTs and real-world evidence.**

38. Mahendraratnam N, Eckert J, Mercon K *et al.* Understanding the need for non-interventional studies using secondary data to generate real-world evidence for regulatory decision making and demonstrating their credibility (2020). https://healthpolicy.duke.edu/sites/default/files/2020-08/Non-Interventional%20Study%20Credibility.pdf

39. Kesselheim AS, Avorn J. The Food and Drug Administration has the legal basis to restrict promotion of flawed comparative effectiveness research. *Health Aff.* 31(10), 2200–2205 (2012).

40. Dreyer NA, Hall M, Christian JB. Modernizing regulatory evidence with trials and real-world studies. *Ther. Innov. Regul. Sci.* 54, 1112–1115 (2020).

Commentary

For reprint orders, please contact: reprints@futuremedicine.com

Journal of Comparative Effectiveness Research

Real-world evidence and nonrandomized data in health technology assessment: using existing methods to address unmeasured confounding?

Cormac J Sammon[1], Thomas P Leahy[1], Sandro Gsteiger[2] & Sreeram Ramagopalan*[,2]

[1]PHMR LTD, London, UK

[2]F. Hoffmann-La Roche Ltd, Basel, Switzerland

*Author for correspondence: sreeram.ramagopalan@roche.com

"Multiple potential approaches have been proposed over this time period, however the majority of them represent variations on a similar approach in which one assesses the impact of a suspected or hypothetical unmeasured confounder on the results observed."

First draft submitted: 16 June 2020; Accepted for publication: 21 July 2020; Published online: 6 August 2020

Keywords: confounding • HTA • real-world evidence

Health Technology Assessment (HTA) bodies are increasingly presented with submissions containing nonrandomized measures of treatment effect, including real-world evidence. The consideration of such evidence has been hampered by uncertainty surrounding the potential impact of unmeasured confounding – often leading to such data being disregarded in the decision-making process. Methods to quantitatively explore the potential impact of unmeasured confounding on estimated treatment effects exist and offer potential to support the use of nonrandomized data in HTA. This article provides an overview of these methods, highlights their underutilization in HTA and considers the steps that would be required to increase their use in this field.

Nonrandomized measures of treatment effect are increasingly being submitted to HTA bodies in the form of unanchored indirect treatment comparisons and real-world evidence (RWE) of comparative effectiveness. It is well known that the key issue with such evidence is the potential for unmeasured confounding to bias the effects observed. That is, as treatment allocation is not randomized, individuals receiving different treatments may systematically differ in ways that impact their risk of the outcome(s) under study, thereby biasing comparisons of these groups. Where these factors have been measured, confounding can be accounted for through appropriate study design and analysis. However real-world data sources are often not fit for purpose in this regard, lacking complete, high-quality measurements of all important confounders. As a result, when faced with these data, HTA bodies commonly provide qualitative descriptions of their concerns regarding the uncertainty unmeasured confounding introduces into the decision-making process and highlight concerns about the extent to which this complicates the interpretation of quantitative assessments of clinical and cost–effectiveness. In many cases the uncertainty raised can contribute to negative decisions regarding reimbursement, thereby impacting patient access to potentially cost-effective treatments.

In a field in which the quantitative synthesis of diverse data to inform decision-making is commonplace, the limited discussion of quantitative methods to explore the issue of unmeasured confounding in HTA submissions is surprising. For example, in their very useful guidelines regarding 'the use of observational data to inform estimates of treatment effectiveness in technology appraisal' and 'methods for population-adjusted indirect comparisons in submissions with NICE' the NICE Decision Support Unit gives very limited advice about how to address unmeasured confounding quantitatively, highlighting this as an area for future research [1,2]. In Germany, IQWIG's methods guidance allows for the consideration of treatment effects from nonrandomized studies where 'dramatic effects' are observed, citing a relative risk of greater than 10 and statistical significance at the 1% level as an effect broadly in a range dramatic enough to be unlikely to be due to unmeasured confounding. However, IQWIG state

Future Medicine

that this is not a rigid threshold and provide little further guidance on this topic [3]. The concept of 'dramatic effects' appears more readily acceptable for decision-making by the German Federal Joint Committee (G-BA) than statistical adjustment methods. An analysis of past dossiers showed that, based on the argument of 'dramatic effects', G-BA accepted a larger proportion of unadjusted (naive) comparisons than adjusted indirect comparisons [4]. The lack of specific guidance on approaches to quantitatively explore unmeasured confounding is not limited to the UK and German guidelines [5–8].

The lack of consideration given to this area is particularly notable given that approaches to quantify the potential impact of hypothetical unmeasured confounders exist and have been under development in the field of (pharmaco)epidemiology for decades. Multiple potential approaches have been proposed over this time period, however the majority of them represent variations on a similar approach in which one assesses the impact of a suspected or hypothetical unmeasured confounder on the results observed [9]. The approaches proposed have differed in several ways. Some used external data from another source to define the strength of a suspected confounder, some assessed the impact of confounders of different strengths on the treatment effect, and still others focused on identifying the minimum strength of a confounder required to adjust an observed treatment effect to the null [9–11]. Analytic approaches have also varied. Some being deterministic, others probabilistic, some implemented in a frequentist, others in a Bayesian framework [9,12,13]. In terms of the types of treatment effects, confounders and outcomes, methods have been developed (among others) for relative risks, odds ratios, risk differences and hypothetical binary and continuous confounders [9,11,13,14]. Similar ideas have also been proposed in other areas of evidence synthesis research, quantifying for example the (hypothetical) level of bias needed to change a treatment recommendation resulting from a network meta-analysis [15].

As an example of how these types of methods work, we use two of the most simple approaches mentioned above to illustrate the potential impact of unmeasured confounding in a nonrandomized study comparing the overall survival of anaplastic lymphoma kinase-positive non-small-cell lung cancer patients who received alectinib with those who received ceritinib [16]. In the study, single-arm trial data on alectinib treated patients were compared against real-world data on ceritinib treated patients from an electronic health record database. A doubly robust approach was used to account for measured confounders resulting in an adjusted hazard ratio of 0.65 (CI: 95% 0.48–0.88). In the discussion, the authors noted that unmeasured confounding may be an issue, for example, due to the absence of complete information on a key prognostic score in the electronic health record database. Since these sensitivity analysis methods are typically applied on the relative risk scale, the first step in applying them is to approximate the adjusted risk ratio (ARR) using the square-root transformation [17]. Applying the transformation to the hazard ratio (HR), we obtain an estimated ARR of 0.74. To then apply the array approach described in [9], one could assume that the prevalence of a hypothetical confounder in the unexposed group is 0.2 (20%), then for varying strengths of association between the hypothetical confounder and the disease outcomes and prevalence of the confounder in the exposed group, a fully adjusted exposure RR can be estimated (Figure 1). Additionally, the E-value as described in [11] can be calculated as, E-value = $ARR^* + \sqrt{ARR^* \times (ARR^* - 1)}$, where $ARR^* = 1/ARR$ since ARR <1. In this example, the E-value is 2.03.This means that to explain away the ARR of 0.74 there would need to be an unmeasured confounder associated with at least 2.03-times the risk of both mortality and alectinib treatment, above and beyond the measured covariates [11]. Notably, one could also carry out the same procedures on the upper or lower bound of the CI [11].

Given the existence of these methods, the question that begs is where the hurdles lie in implementing them in existing HTA frameworks? Programmatically, operationalizing the tools should not be a major hurdle. Those HTA bodies whose decisions are based on relative effectiveness assessments could utilize one of the existing tools discussed above to pressure-test the nonrandomized measures of relative effect presented to them to an extent they are comfortable with. Those HTA bodies that use cost–effectiveness frameworks could use a similar approach or could potentially look to build the probabilistic or Bayesian sensitivity analysis approaches, mentioned above, directly into the probabilistic sensitivity analysis already found in most cost–effectiveness models submitted to HTA bodies. This would allow for the uncertainty due to unmeasured confounding to be captured alongside all of the other sources of uncertainty thereby fitting with the current decision-making framework. Some additional work may be required to ensure the outcome types typically encountered in HTA submissions can be adjusted using the appropriate methods but given the stage of development of the methodological field, one would not expect this to be an issue.

It appears that the bigger hurdle may be in setting out the framework required to govern how to parametrize the chosen tools, that is, providing answers to questions along the lines of how big of a confounder can be considered

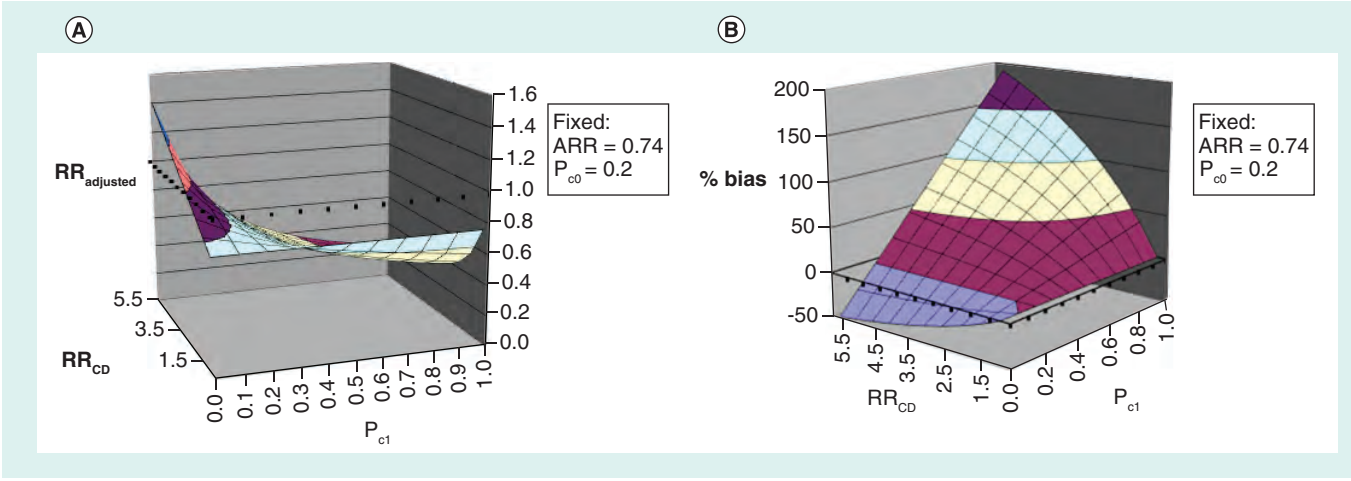

Figure 1. Fully adjusted relative risk ($RR_{adjusted}$) surface (A) and percentage of bias surface (B) as a function of the strength of association between a confounder and mortality (RR_{CD}) and the prevalence of the confounder in the alectinib treated group PC_1 using the array approach.

'unrealistic' or 'unlikely'? Should this be determined on an assessment-specific basis, an indication-specific basis or overall for all submissions? Should there be an onus on manufacturers to capture data from external sources in order to better inform the parameters of these analyses? Should there be an onus on the HTA bodies side to have their clinical experts determine the parameters of these sensitivity analyses? If the latter, how should expert elicitation best be carried out in practice?

Answering these questions will require careful thought on the part of multiple stakeholders and likely some sort of consultation process, however we believe that any reimbursement body committed to utilizing single-arm data and/or RWE to accelerate patient access to therapy needs to find a way to align on these questions. This may require a programme of work focused on reviewing and publicizing the tools available to quantitatively explore unmeasured confounding to the HTA audience, as we seek to begin to do with this article, and carrying out any additional methodological and conceptual work required to allow for their incorporation into decision-making frameworks. Given recent initiatives by a number of HTA bodies to better consider the potential use of RWE for HTA, now may be a pertinent time for greater focus on this area [18–22].

Financial & competing interests disclosure

This work was funded by FH-La Roche Ltd; S Ramagopalan and S Gsteiger are employees of FH-La Roche Ltd. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

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

References

1. Faria R, Hernadez Alava M, Manca A, Wailoo A. The use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data. (2015). NICE Decision Support Unit, Technical Support Document 17. http://nicedsu.org.uk/technical-support-documents/observational-data-tsd/
2. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med. Decis. Making* 38(2), 200–211 (2018).
3. Institute for Quality and Efficiency in Health Care. General methods (2015). https://www.iqwig.de/en/methods/methods-paper.3020.html
4. Eichinger B, Ecker C. Indirekte Vergleiche in der frühen Nutzenbewertung in Deutschland–Erfahrungen aus den ersten 5 Jahren. *Gesundh. Ökon Qual. Manag.* 21(06), 279–287 (2016).

5. Health Information and Quality Authority. Guidelines for evaluating the clinical effectiveness of health technologies in Ireland (2019). https://www.hiqa.ie/reports-and-publications/health-technology-assessment/guidelines-evaluating-clinical-effectiveness
6. Wells G, Sultan S, Chen L. Indirect evidence: indirect treatment comparisons in meta-analysis. The Canadian agency for drugs and technologies in health (CADTH) report (2009). https://www.cadth.ca/sites/default/files/pdf/H0462_itc_tr_e.pdf
7. Institute for Clinical and Economic Review. 2020–2023 Value Assessment Framework (2020). https://icer-review.org/wp-content/uploads/2019/05/ICER_2020_2023_VAF_013120-1.pdf
8. Australian Department of Health. Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee (2016). https://pbac.pbs.gov.au/content/information/files/pbac-guidelines-version-5.pdf
9. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol. Drug Saf.* 15(5), 291–303 (2006).
10. Stürmer T, Glynn RJ, Rothman KJ, Avorn J, Schneeweiss S. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. *Med. Care* 45(Suppl. 10), S158 (2007).
11. Vanderweele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann. Intern. Med.* 67(4), 268–274 (2017).
12. Barrowman MA, Peek N, Lambie M, Martin GP, Sperrin M. How unmeasured confounding in a competing risks setting can affect treatment effect estimates in observational studies. *BMC Med. Res. Methodol.* 19(1), 166 (2019).
13. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data.* Springer Publishing Company, Inc, NY, USA (2009).
14. Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* 22(1), 42–52 (2011).
15. Phillippo DM, Dias S, Ades A, Didelez V, Welton NJ. Sensitivity of treatment recommendations to bias in network meta-analysis. *J. R. Stat. Soc. Ser. A Stat. Soc.* 181(3), 843–867 (2018).
16. Davies J, Martinec M, Delmar P *et al.* Comparative effectiveness from a single-arm trial and real-world data: alectinib versus ceritinib. *J. Comp. Eff. Res.* 7(9), 855–865 (2018).
17. Vanderweele T. On a square-root transformation of the odds ratio for a common outcome. *Epidemiology* 28(6), e58 (2017).
18. Murphy G. Use of real-world evidence in single-drug assessments. CADTH, Ottawa, Canada (2018). https://www.cadth.ca/use-real-world-evidence-single-drug-assessments-environmental-scan
19. NICE. Increasing use of health and social care data in guidance development. https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-guidelines/how-we-develop-nice-guidelines/data-and-analytics-statement-of-intent
20. Institute for Quality and Efficiency in Health Care. Concepts for the generation of routine practice data and their analysis for the benefit assessment of drugs according to §35a Social Code Book V (2020). https://www.iqwig.de/download/A19-43_Routine-practice-data-for-the-benefit-assessment-of-drugs_Rapid-report_V1-0.pdf
21. Hampson G, Towse A, Dreitlein WB, Henshall C, Pearson SD. Real-world evidence for coverage decisions: opportunities and challenges. *J. Comp. Eff. Res.* 7(12), 1133–1143 (2018).
22. French National Authority for Health. Innovative medicines assessment action plan (2020). https://www.has-sante.fr/upload/docs/application/pdf/2020-03/innovative_medicine_action_plan_27.01.20.pdf

Commentary

For reprint orders, please contact: reprints@futuremedicine.com

Journal of Comparative Effectiveness Research

Life after COVID-19: R WE going to help?

Sreeram V Ramagopalan*,[1] & Radek Wasiak[2]

[1]Global Access, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, CH-4070, Basel, Switzerland

[2]Cytel Ltd, Hamilton House, Mabledon Place, London WC1H 9BB, UK

*Author for correspondence: sreeram.ramagopalan@roche.com

"Less fragmented data can help in all aspects of the patient journey, with the ultimate goal of improving patient (and caregiver) holistic well-being (physically, mentally and financially) and concomitantly reducing burden and cost to healthcare systems."

First draft submitted: 9 April 2020; Accepted for publication: 14 April 2020; Published online: 11 May 2020

Keywords: coronavirus • COVID-19 • real-world evidence

The threat of COVID-19 has led to an unprecedented level of cooperation among academia, industry, policy makers and public health officials to find ways to diagnose, prevent and cure the condition. Although key work is interventional trials testing vaccines and treatments, observational studies (or real-world evidence [RWE]) have proved valuable in aiding decision-making [1–4]. Awareness of how RWE is generated, its limitations and the biases presented that need to be addressed using statistical methods to gain meaningful insights, is now commonplace and should not impede its use and acceptance [5]. The research activities related to COVID-19 should hopefully help end the often unproductive debate about the superiority of randomized clinical trials over RWE [6].

Analogizing to the 2008 global financial crisis, which negatively impacted individual health and access to care and forced changes to service provision in budget constrained healthcare systems [7,8], we will likely observe material changes after the pandemic is over. What is different about this shock to our lives and economies is the substantially greater existence of data and the ability to process this in vast quantities.

Data-enabled decision-making has the potential to offer significant benefits to patients, health systems and society as a whole; however, in order to fully capitalize on this, more data need to be made available. Imagine if full medical records were available on people worldwide, in real time, potentially also linked to their social media and wearable technology accounts. If used for research purposes (complying with all privacy and legal guidance) this could have meaningfully assisted efforts to build predictive models for COVID-19 and associated complications [9] and to aid flattening of the COVID-19 transmission curve, translating to extensive global health and economic gains.

How could more data help after the COVID-19 pandemic? Cancer is a highly relevant example where there are high unmet needs and limited data available. A survey of over 4000 cancer patients in more than ten countries by All. Can [10], found that there were often delays in disease diagnosis, inefficiencies in cancer management and that the disease caused considerable financial challenges for sufferers through loss of earnings. It is likely that these findings can be extrapolated to some extent to other diseases. Less fragmented data can help in all aspects of the patient journey, with the ultimate goal of improving patient (and caregiver) holistic well-being (physically, mentally and financially) and concomitantly reducing burden and cost to healthcare systems.

New methods can help convert data to insight; artificial intelligence techniques have already been shown to assist on cancer image data to improve diagnosis [11]. With prospective validation of these techniques with imaging data directly from the clinic, deployment of these systems could reduce strain on the healthcare workforce and ensure high-risk patients are rapidly followed up. Although precision medicine is currently most applicable to oncology, where genomically targeted therapy can lead to better clinical outcomes [12], medicine can become more personalized and more efficient with more complete and less fragmented data. Actionable information such as quality of life data are important and whether this is tracked using wearables, social media or via self-monitoring apps [13,14], it can empower both patients and healthcare practitioners to address any deterioration in symptoms that otherwise may escape detection in between routine clinical visits. Improvement of patient outcomes potentially

Future Medicine

not only enables them (and their caregivers) to carry on working, which has a positive societal impact, it also lessens healthcare resource use.

The need to develop new medicines remains, despite most attention being placed on healthcare interventions to support ending the COVID-19 pandemic. Real world data can provide supportive information to help with designing (e.g., defining appropriate [potentially digital and more patient-centric] end points to use) and interpreting (e.g., understanding generalizability of trial populations) randomized clinical trials. Real world data can also be used to act as a comparator when it is difficult or unethical to run a placebo-controlled study. All of these efforts can assist in bringing patients faster access to medicines and ultimately improving their lives and bettering society in general. While the health and financial advantages of more data availability are tangible, there are nevertheless data privacy, quality, fragmentation, infrastructure and logistical challenges that need to be surmounted [15–17]. However, COVID-19 has shown that when a crisis hits, we all can work together and commit to a greater good. Perhaps this is an opportunity that the pandemic is affording us: putting things in place now to improve health and wealth outcomes for all and limit future healthcare emergencies.

Financial & competing interests disclosure

SV Ramagopalan is an employee of Roche. R Wasiak is an employee of Cytel. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

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

References

1. Verity R, Okell LC, Dorigatti I *et al.* Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect. Dis.* doi:10.1016/S1473-3099(20)30243-7 (2020) (Epub ahead of print).
2. Kraemer MUG, Yang C-H, Gutierrez B *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* doi:10.1126/science.abb4218 (2020) (Epub ahead of print).
3. He X, Lau EHY, Wu P *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* doi:10.1038/s41591-020-0869-5 (2020) (Epub ahead of print).
4. Leung NHL, Chu DKW, Shiu EYC *et al.* Respiratory virus shedding in exhaled breath and efficacy of face masks. *Nat. Med.* doi:10.1038/s41591-020-0843-2 (2020) (Epub ahead of print).
5. Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. Avoidable flaws in observational analyses: an application to statins and cancer. *Nat. Med.* 25(10), 1601–1606 (2019).
6. Collins R, Bowman L, Landray M, Peto R. The magic of randomization versus the myth of real-world evidence. *N. Engl. J. Med.* 382(7), 674–678 (2020).
7. Karanikolos M, Heino P, McKee M, Stuckler D, Legido-Quigley H. Effects of the global financial crisis on health in high-income OECD countries: a narrative review. *Int. J. Health Serv. Plan. Adm. Eval.* 46(2), 208–240 (2016).
8. Antonanzas F. The impact of the economic downturn on healthcare in Spain: consequences and alternatives. *Expert Rev. Pharmacoecon. Outcomes Res.* 13(4), 433–439 (2013).
9. Wynants L, Van Calster B, Bonten MMJ *et al.* Systematic review and critical appraisal of prediction models for diagnosis and prognosis of COVID-19 infection (2020). www.medrxiv.org/content/10.1101/2020.03.24.20041020v2
10. All. Can patient survey (2019). www.all-can.org/what-we-do/research/patient-survey/
11. McKinney SM, Sieniek M, Godbole V *et al.* International evaluation of an AI system for breast cancer screening. *Nature* 577(7788), 89–94 (2020).
12. Singal G, Miller PG, Agarwala V *et al.* Association of patient characteristics and tumor genomics with clinical outcomes among patients with non-small cell lung cancer using a clinicogenomic database. *JAMA* 321(14), 1391–1399 (2019).
13. McDonald L, Malcolm B, Ramagopalan S, Syrad H. Real-world data and the patient perspective: the promise of social media? *BMC Med.* 17(1), 11 (2019).
14. Denis F, Basch E, Septans A-L *et al.* Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. *JAMA* 321(3), 306–307 (2019).
15. McDonald L, Ramagopalan S. Do we need to care about the closure of care.data? (2016). https://blogs.biomedcentral.com/on-medicine/2016/08/19/need-care-closure-care-data/
16. Graham S, McDonald L, Wasiak R, Lees M, Ramagopalan S. Time to really share real-world data? *F1000Research* 7, 1054 (2018).
17. European data strategy. Making the EU a role model for a society empowered by data (2020). https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

White Paper

Journal of Comparative Effectiveness Research

Strengthening pharma's contract with society: the value of trusted partnerships between pharma and healthcare facilitated by real-world data

Michael J Seewald*[,1], Jonathan M Plumb[2], Ben Gutierrez[3], Johan Liwing[4], Maurille Feudjo Tepie[5], Mette Hammer[6], Robert LoCasale[7], Sajan Khosla[8], Donald Yin[9], Chris L Pashos[10] & Mark Nagy[11]

[1]Novartis Pharma AG/Pharma Medical Affairs, Basel, Switzerland
[2]Ferring Pharmaceuticals SA, St Prex, Switzerland
[3]GSK, Philadelphia, PA, USA
[4]Janssen, Stockholm, Sweden
[5]Center for Observational Research, Amgen, Cambridge, UK
[6]Novo Nordisk, Copenhagen, Denmark
[7]Sanofi, Bridgewater, NJ, USA
[8]Real-World Evidence Center of Excellence, AstraZeneca, Cambridge, UK
[9]Merck, Kenilworth, NJ, USA
[10]AbbVie, Chicago, IL, USA
[11]Lilly, Indianapolis, IN, USA
*Author for correspondence: Tel.: +41 79 543 2073; mseewald@gmail.com

This White Paper is authored by 11 industry real-world evidence (RWE) experts, with support from IQVIA, as part of the 'RWE Leadership Forum': a group of industry leaders who come together as noncompetitive partners to understand and respond to internal or external RWD/E challenges and opportunities with a single expert voice. Herein we aim to clarify the rules of engagement between pharma and healthcare in order to establish trust-based partnerships, which will unlock unique value for society, including the medical community and the ultimate beneficiary, the patient.

First draft submitted: 5 December 2019; Accepted for publication: 13 December 2019; Published online: 22 January 2020

Keywords: healthcare • observational research • partnerships • pharma • real-world data • real-world evidence

A rare opportunity

As health systems find themselves under increasing pressure, the pharmaceutical industry has a unique opportunity to broaden the very basis of their engagement with health systems beyond predominantly one-dimensional relationships focused on access to innovative treatments. Instead, the pharmaceutical industry can look to build trusted partnerships around a common purpose of supporting the delivery of better patient outcomes through application of data-driven insights to improve treatment and care pathways, to advance medical practice and to optimize the utilization of healthcare resources.

The term real world data (RWD) describes data that are routinely collected within health systems on a patient's health status, in a purely observational fashion. RWD has the potential to play a critical role in establishing trusted partnerships: its very origin from within health systems, makes RWD particularly suited for generating evidence in a transparent and objective way through collaborations. As such, the generation of RWD can help foster trust and support new research partnerships that build a shared understanding around the intent of collecting data, the purpose this will serve, which trade-off decisions the data will inform and the value such partnerships offer to each party.

Future Medicine

The emergence of interdependencies

The traditional relationships between the pharmaceutical industry and healthcare systems are being fundamentally re-defined, by moving beyond typical transactional roles as seller and buyer, respectively.

As health systems struggle under increasing financial pressures, engagement on the sole basis of providing innovative treatments for payment makes it difficult for both parties to reconcile conflicting priorities. Establishing long-term, trust-based partnerships around broad, shared objectives such as aligning treatment around optimal patient outcomes, enables health systems and the pharmaceutical industry to solve problems jointly.

Health systems and the pharmaceutical industry recognize the value of evidence and insight generated by RWD: alongside governments, the pharmaceutical industry has fueled an increasingly rich RWD landscape, as some health systems lack the expertise or resources to generate insight from those data assets themselves.

Research-focused partnerships transcend the traditional monetization of data and aim to solve the challenge of sustainable innovation for better patient outcomes. However, in order to build these partnerships, trust must be established between both parties.

The value of trusted partnerships

The defining feature of trusted partnerships is the long-term pursuit of mutually beneficial objectives, in a transparent way, by combining complementary assets, resources and capabilities. As the home of deep scientific, clinical and analytical expertise, with an ability to draw on significant resources, the pharmaceutical industry has a lot to contribute to establish trust-based partnerships. In the case of trusted partnerships between health systems and the pharmaceutical industry, this common ground can be found at the intersection of innovation, efficiency and patient outcomes, with RWD facilitating transparency, objectivity and stakeholder engagement:

- Patient outcomes: achieving better patient outcomes in the real world is the ultimate prize for trusted partnerships between health systems and the pharmaceutical industry. At the very heart of this lies the optimization of treatment and care pathways;
- Innovation: long-term collaborations as part of a network of relationships, instead of one-off studies, can facilitate the build out of novel RWD infrastructure, which drives unique insights to advance medical practice;
- Efficiency: improving the operational performance of healthcare systems is a major driver of mutual benefit in trusted partnerships. Efficiency gains help to ensure the sustainable delivery of high-quality care and unlock funds for broadening patient access to cutting edge innovation.

However, there are important considerations for defining the common ground between health systems and the pharmaceutical industry, specifically, the issue of scale of such partnerships. We believe two complimentary modes of working should be considered:

- Scaling up from local pilots: establishing exemplars of best practice, by demonstrating the art of the possible locally first, for example, by partnering directly with one to two centers, and then transferring learnings across the wider health system;
- Driving structural change at (inter-)national level: partnering with national and international bodies, including regulators and payers, to drive change, for example, by furthering methodology and data standards.

Crucially, to reap the full benefits from optimizing along the dimensions of innovation, efficiency and patient outcomes, learning healthcare systems [1] need to embrace a virtuous cycle of continuous improvement and be open to international best practice sharing. While healthcare systems operate predominantly locally, the pharmaceutical industry with its global reach is well placed to facilitate such exchange of learnings across countries.

Implications for partnership models: from transactional to relationship building

Historically, many collaborations between healthcare systems and the pharmaceutical industry tended to be transient in nature and only on a need-to basis, with short-term funding and engagement commitments, and therefore they do not always support a long-term and sustainable, mutually beneficial partnership model. Short-term focus represents a major barrier to unlocking the unique value of novel RWD assets. Lead times of 5 years or more are common for such novel RWD assets to reach the requisite data volumes and levels of data maturity needed for high-quality evidence generation, well beyond the typical initial 2-year sourcing and curation period to secure data access; hence,

the importance of establishing long-term relationships that enable the sustainable collection of RWD by healthcare systems.

To move to trust-based partnerships, new models are required that ensure long-term commitment, transparency and joint ownership of insights, all in full compliance with data privacy requirements and ethical standards. Recognizing the need for combining the best available data from all partners to solve a common problem is a key factor for successful collaborations, such as public–private partnership projects [2,3], multi-stakeholder collaborations [4,5] or single company-led, multi-stakeholder collaborations [6–8].

Requirements for successful trusted partnership models

For trust-based partnership models to succeed, several barriers need to be overcome which requires a concerted, multi-stakeholder effort. Specifically, the following agreements and clarifications are needed from healthcare systems:

- Alignment on the rules and models for access and contribution to data:
  - Clarification of the value of improving the quality of healthcare data in the pursuit of delivering better patient outcomes, and communicate said value to patients in order to encourage sharing of their data for research purposes;
  - Clarification of 'research' definition to facilitate data sharing under preferential General Data Protection Regulation (GDPR) or Health Insurance Portability and Accountability Act (HIPAA) requirements;
  - Streamlined approval processes, for example, pre-agreed study archetypes, standard operating procedures (SOPs) for standard research approaches;
  - Accelerated approval of data linkage.
- Alignment on fit-for-purpose methodologies and suitability of data sources:
  - Clarification of how to control for bias, beyond solely relying on randomization;
  - Clarification of suitable data sources for assessing clinical effectiveness and safety, especially when relying on disease-specific intermediate outcomes rather than typically well documented final outcomes such as death.
- Alignment on governance and transparency, as pre-requisites for data access and patient advocacy:
  - Agreement on ethical redlines for research;
  - Agreement on the principles of assignment of responsibilities;
  - Agreement on ownership and acceptable use of deliverables, including publication rights.

Health systems, and patients, have much to gain from adopting those changes in order to facilitate trust-based partnerships with the pharmaceutical industry and to reap the full benefits they have to offer.

Commitment by the pharmaceutical industry

To create momentum in advancing the trusted partnership agenda, the authors of this commentary endorse the following principles. They have been designed to ensure fairness, transparency and, ultimately, maximize the value of such partnerships to healthcare.

- Responsible use of health data: we believe that health data should be handled responsibly, in adherence to data privacy and ethical standards. Where laws and regulations allow the re-use of health data for research purposes without specific consent, (for example, to promote public health), we believe that industry, academia and legislators have the joint responsibility to educate individuals and society about the significant benefits of appropriate data re-use. Individuals should have transparency over the use of their personal health data, therefore we are committed to publication of all studies evaluating the effectiveness and safety of interventions in public registries and timely publication of study results, ideally in peer-reviewed scientific journals;
- Balancing stakeholder interests: we believe that the digital age will transform many aspects of biomedical research, clinical practice and healthcare systems. To harvest the full benefit of these developments for patients and society, a balance must be struck between the individual's privacy rights, interests and autonomy over the use of their personal health data and the broader societal gains in obtaining new knowledge and advancing medicines. As perceptions on aspects of these debates vary broadly not only from country to country, but also between generations, this requires a long-term, inclusive societal debate. We are committed to be a transparent, pro-active stakeholder in this discourse, and we will publicly share our views on key aspects of the discussion;

- Multi-stakeholder collaborations: we believe multi-stakeholder collaborations, which aim to develop and implement responsible data sharing approaches, are an efficient, transparent and responsible way to facilitate access to (novel) health data for research and healthcare provision. These initiatives rely on a fit-for-purpose data model designed to ensure adequate data privacy and security, coupled with broad representation from different stakeholders and an independent Steering Committee that ensures strong governance and ethical use of health data. We are committed to support the adoption of such multi-stakeholder initiatives and welcome collaborations with a wide range of potential partners across multiple disease areas;
- Collaborative cross-industry approach: we believe the pharmaceutical industry should collaborate on nonproduct, disease area-specific evidence generation, to further expand on the foundations laid by evidence-based medicine [9] and contribute to the acceleration of medical practice. We commit to share the insights derived from RWD using the appropriate channels, to the extent permitted under the data license agreements, in order to improve the collective disease area understanding and empower patients to seek the right care, help healthcare professionals best treat their patients and support healthcare systems in preparing for future challenges;
- Long-term commitment: wherever possible and meaningful, we believe in supporting long-term commitments to the sustainable creation and curation of new data assets, including the routine capture of outcomes data, embedded within healthcare systems. Such long-term commitments ensure that innovation continues to live on and that value is shared with the very sources of the data collection effort, instead of being exclusively captured by the sponsor(s). The success [10] of such efforts is dependent on collaboration between multiple stakeholders, including electronic medical record providers;
- Frameworks and methodological standards: we support existing good practice for the generation of high-quality RWE, including ISPOR-ISPE guidance, for example, "*Good Practices for Real-World Data Studies of Treatment and/or Comparative Effectiveness*" [11]; "*Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies*" [12] and "*Use of Electronic Health Record Data in Clinical Investigations*" [13]. Furthermore, we welcome collaborations with regulators and commit to engage in constructive dialogue on shaping frameworks for advancing the generation and acceptability of RWE in support of stakeholder decisions.

By endorsing these principles, we are paving the way for the pharmaceutical industry to place itself even further at the forefront as a trusted partner in delivering better outcomes for patients.

This document lays the foundations for establishing trusted partnerships between the pharmaceutical industry and health systems. Naturally, several practical questions remain such as how to ensure equitable investments in healthcare are made, where long-term value is not solely realized in a small number of specialist centers, but widely shared across the system, and what is the most appropriate mechanism with which to make an investment? Questions such as these deserve their own dedicated discussion in future papers.

The healthcare environment is ripe for embracing trust-based partnerships. Indeed, the long-term sustainability of a vibrant learning healthcare system depends on establishing trust-based partnerships which also involve the patient and medical communities. The pharmaceutical industry now has a rare opportunity to fundamentally re-define their relationships with health systems, healthcare providers and patients, with focus on reconciling innovation, efficiency and patient outcomes, with RWD in a facilitating role as the source of insight, transparency and objectivity.

By ensuring healthcare decisions, including both pharmacological and nonpharmacological interventions, are consistently informed by a growing body of high-quality evidence, those trust-based partnerships will fulfill their promise and unlock unique value for society, with benefits delivered to many stakeholders, including the medical community and the ultimate beneficiary, the patient.

Financial & competing interests disclosure

The authors are employees of pharmaceutical companies. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Writing support was provided by IQVIA.

References

1. Friedman C, Rubin J, Brown J *et al.* Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *J. Am. Med. Inform. Assoc.* 22(1), 43–50 (2014).
2. Vaudano E. Public–private partnerships as enablers of progress in the fight against mental disorders: the example of the European Innovative Medicines Initiative. *Eur. Psychiatry* 50, 57–59 (2018).
3. Harmony-alliance.eu Algorithms To Transform Knowledge Into Better Medicine Outcomes. (2019). www.harmony-alliance.eu/en/news/wp7/algorithms-to-transform-knowledge-into-better-medicine-outcomes
4. Hripcsak G, Duke JD, Shah NH *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inform.* 216, 574–578 (2015).
5. Miller R, Wong J. Using oncology real-world evidence for quality improvement and discovery: the case for ASCO's CancerLinQ. *Future Oncol.* 14(1), 5–8 (2018).
6. Durand-Zaleski I, Manley-Daumont M, Baas P *et al.* I-O Optimise: developing a unique multinational real-world evidence-based research platform in oncology. *Value Health* 20(9), A742–A743 (2017).
7. Garside J, Healy N, Besson H *et al.* PHEDRA: using real-world data to analyze treatment patterns and ibrutinib effectiveness in hematological malignancies. *J. Comp. Eff. Res.* 7(1), 29–38 (2018).
8. Catalyzing the Transformation of Healthcare Innovation (2019). https://newdigs.mit.edu/catalyzing-transformation-healthcare-innovation
9. A Brief History of Evidence-based Practice – Evidence Based Practice in Optometry EBP Australia UNSW (2019). www.eboptometry.com/content/optometry/article/brief-history-evidence-based-practice-0
10. Notice of Proposed Rulemaking to Improve the Interoperability of Health Information (2019). www.healthit.gov/topic/laws-regulation-and-policy/notice-proposed-rulemaking-improve-interoperability-health
11 . Berger M, Sox H, Willke R *et al.* Recommendations for Good Procedural Practices for Real-World Data Studies of Treatment Effectiveness and/or Comparative Effectiveness Designed to Inform Health Care Decisions: Report of the Joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making. *Value Health* 20(8), 1003–1008 (2017).
12 . Wang S, Schneeweiss S, Berger M *et al.* Reporting to improve reproducibility and facilitate validity assessment for healthcare Database Studies V1.0. *Value Health* 20(8), 1009–1022 (2017).
13. US Department of Health and Human Services Food and Drug Administration. Use of Electronic Health Record Data in Clinical Investigations, Guidance for Industry (2018). www.fda.gov/regulatory-information/search-fda-guidance-documents/use-electronic-health-record-data-clinical-investigations-guidance-industry

Review

For reprint orders, please contact: reprints@futuremedicine.com

Journal of Comparative Effectiveness Research

Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants

Victoria Allan*[,1], Sreeram V Ramagopalan[1], Jack Mardekian[2], Aaron Jenkins[3], Xiaoyan Li[4], Xianying Pan[5] & Xuemei Luo[6]

[1]Centre for Observational Research & Data Sciences, Bristol-Myers Squibb, Uxbridge, UK
[2]Statistics, Global Biometrics and Data Management, Pfizer Inc., New York City, NY, USA
[3]Patient Health & Impact, Outcomes & Evidence, Pfizer Ltd, Tadworth, UK
[4]Worldwide Health Economics and Outcomes Research, Bristol-Myers Squibb, Lawrenceville, NJ, USA
[5]Pharmacoepidemiology, Bristol-Myers Squibb, Lawrenceville, NJ, USA
[6]Patient Health & Impact, Outcomes & Evidence, Pfizer Inc., Groton, CT, USA
*Author for correspondence: Tel.: +44 189 552 3382; victoria.allan@bms.com

After decades of warfarin being the only oral anticoagulant (OAC) widely available for stroke prevention in atrial fibrillation, four direct OACs (apixaban, dabigatran, edoxaban and rivaroxaban) were approved after demonstrating noninferior efficacy and safety versus warfarin in randomized controlled trials. Comparative effectiveness research of OACs based on real-world data provides complementary information to randomized controlled trials. Propensity score matching and inverse probability of treatment weighting are increasingly popular methods used to address confounding by indication potentially arising in comparative effectiveness research due to a lack of randomization in treatment assignment. This review describes the fundamentals of propensity score matching and inverse probability of treatment weighting, appraises differences between them and presents applied examples to elevate understanding of these methods within the atrial fibrillation field.

First draft submitted: 27 January 2020; Accepted for publication: 3 March 2020; Published online: 18 March 2020

Keywords: atrial fibrillation • comparative effectiveness research • confounding by indication • inverse probability of treatment weighting • oral anticoagulants • propensity score matching

Atrial fibrillation (AF) is a cardiac arrhythmia affecting millions worldwide. In 2010, the global prevalence of AF was estimated at 33.5 million [1] and projections indicate a doubling in the number of patients by 2050 [2], with the ageing population a key contributing factor. AF significantly increases the risk of thromboembolic stroke with patients five-times more likely to experience stroke compared with those without AF [3]. Oral anticoagulants (OACs) are effective in reducing the risk of stroke by 64%, according to a meta-analysis [4]. After decades of warfarin being the only OAC widely available, four direct OACs (DOACs; apixaban, dabigatran, edoxaban and rivaroxaban) were approved after demonstrating noninferior efficacy and safety versus warfarin in randomized controlled trials (RCTs) [5–8].

RCTs evaluate the efficacy and safety of OACs in well-controlled environments with precisely defined study population inclusion and exclusion criteria. RCTs are purposefully designed to minimize sources of bias and provide the ideal conditions for testing whether a cause–effect relationship exists between a treatment and an outcome [9]. Real-world evidence (RWE), derived from data collected in the normal delivery of clinical care, provides complementary information to that obtained from RCTs and is of increasing interest to healthcare decision makers [10]. RWE can be used to establish whether trial efficacy translates into real-world effectiveness, or to explore how a treatment performs in more diverse patient populations with differing levels of adherence

Future Medicine

and compliance, over a longer period of follow-up and in comparative effectiveness research (CER) versus other therapies that have not been examined under trial conditions [11].

The introduction of DOACs has significantly expanded the number of treatment options available to AF patients, yet to date there have been no RCTs comparing the efficacy of one DOAC against another. In the absence of head-to-head RCTs, the comparative effectiveness of DOACs continues to be extensively researched using RWE studies. In a recent systematic review and meta-analysis, Li and colleagues identified 15 RWE studies comparing rivaroxaban versus dabigatran, nine studies comparing rivaroxaban versus apixaban and nine studies comparing apixaban versus dabigatran; there were no comparisons with edoxaban, reflecting its later approval date. The meta-analyzed results showed no significant differences between the three DOACs compared in terms of stroke or systemic embolism risk, whereas apixaban was associated with a significantly lower risk of major bleeding compared with rivaroxaban and dabigatran, and rivaroxaban conferred a significantly higher risk of major bleeding compared with dabigatran. Li and colleagues concluded that the study results may help to guide treatment decisions on the choice of DOACs for patients with AF. However, highlighting a key limitation, the authors reflected that DOAC treatments were not prescribed at random, therefore there could be confounding by indication bias influencing the findings. The potential for bias among the included studies was assessed using the Cochrane Collaboration ROBINS-I (Risk of Bias In Non-randomized Studies of Interventions) evaluation tool, among which 'confounding' and 'selection of participants' are two key assessment domains [12,13].

Confounding by indication may arise in real-world comparative effectiveness studies when there are inherent differences in the patients prescribed the two or more treatments being compared. As defined by Joseph and colleagues, confounding by indication refers to bias in relationship between a treatment and the intended outcome of the treatment due to the clinical reasons for the treatment. The indication for treatment is based upon both the physician's and the patient's judgement of the disease severity, prognosis and expected therapeutic effect of the treatment [14]. RCTs remain the gold standard study design for evaluating the relative benefits and/or risks of treatments, because the process of randomization minimizes imbalances in both observed and unobserved factors that could introduce bias into the assignment of patients to the treatment groups being compared [15]. In an RCT comparing two treatments, patients are equally likely (i.e., have a 50% chance) to receive either treatment, therefore this maximizes the probability that any resulting differences in outcomes between groups are truly due to differences in the treatments and are not due to any observed or unobserved differences in the patients receiving each treatment. CER using real-world data therefore requires rigorous statistical methods, often extending beyond multivariable regression models, to account for the lack of randomization in treatment assignment and control for confounding by indication [16]. Unlike an RCT, these methods are however restricted to addressing only observed confounders (i.e., factors that are measured and collected in the study).

Propensity score matching (PSM) and inverse probability of treatment weighting (IPTW) are increasingly popular methods used to address confounding by indication in RWE studies. Within the AF field, the number of research publications referencing these methods has been increasing year-on-year, as indexed in the PubMed database (Figure 1). Researchers have traditionally relied upon multivariable regression models to adjust for differences in patient characteristics, however, they are now turning to PSM and IPTW because of the stronger theoretical and statistical basis for these methods that has been argued in the literature [17]. While, for the most part, multivariable regression and propensity score-based methods (such as PSM and IPTW) have been found to lead to similar study conclusions, [18,19] from a theoretical standpoint, PSM and IPTW aim to achieve a balanced distribution of confounders across treatment groups and thereby more closely emulate the properties of an RCT [20]. From a statistical standpoint, propensity score-based methods have been shown to lead to more robust and less biased estimations of the treatment effect when there are few outcome events relative to the number of potential confounders (i.e., fewer than eight events per confounder) [21]. In Li and colleagues' systematic review of 15 real-world studies reporting comparisons between DOACs, two thirds of studies continued to use multivariable logistic or survival regression models alone to address confounding, with one third of studies opting for more robust PSM and IPTW methods. The studies using PSM and IPTW methods were more often assessed as being at low risk of bias [12].

While PSM and IPTW endeavor to achieve the same objective in balancing out differences between treatment groups, the two methods provide a different measurement of the treatment effect and this should be interpreted accordingly. When applied to the same data, PSM and IPTW may not always point to the same findings suggesting that these methods are not strictly interchangeable. For example, a study comparing dabigatran versus warfarin among real-world AF patients reported a hazard ratio of 0.77 (95% confidence interval: 0.54–1.09) for the risk of stroke and 0.75 (0.65–0.87) for the risk of major bleeding when estimated using PSM, contrasting with 0.00 (0.00–

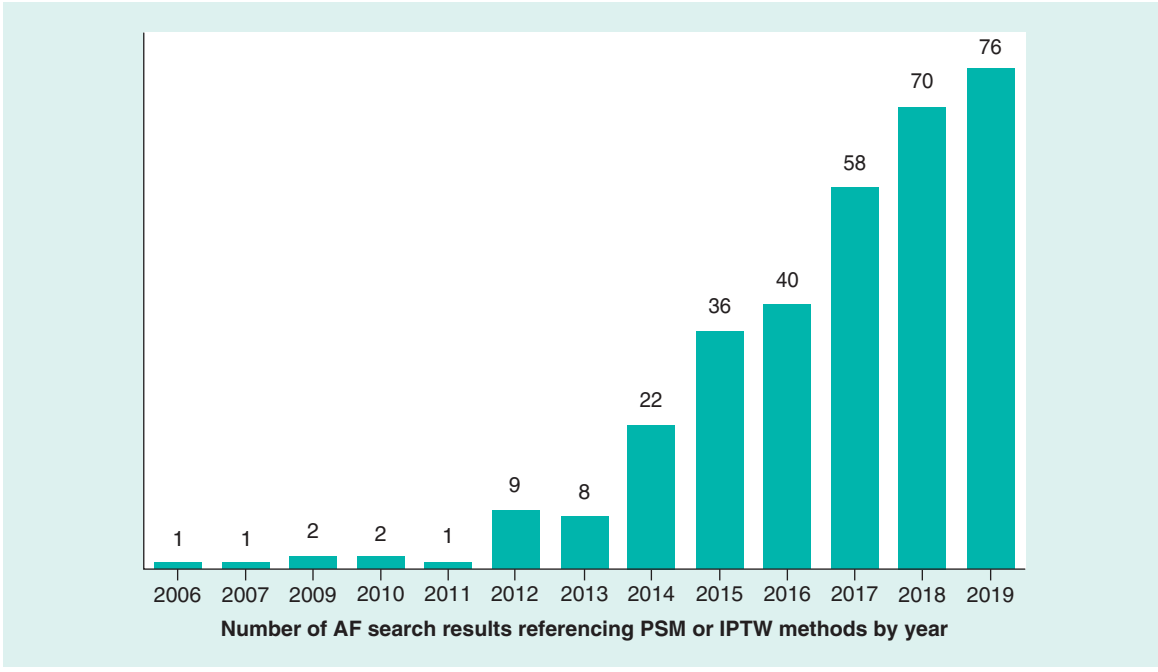

Figure 1. Increase in propensity score matching and inverse probability of treatment weighting methods within comparative effectiveness research of oral anticoagulants (2006–2019). Search results per year were downloaded from the PubMed database (https://www.ncbi.nlm.nih.gov/pubmed/) on 29-Nov-2019 based on the following search terms: (anticoagulant* OR warfarin OR apixaban OR dabigatran OR rivaroxaban OR edoxaban) AND (atrial fibrillation) AND (propensity score OR inverse probability of treatment weighting).
AF: Atrial fibrillation; IPTW: Inverse probability of treatment weighting; PSM: Propensity score matching.

0.56) and 0.08 (0.08–0.10), respectively, when estimated using IPTW [22]. As shown, PSM and IPTW procedures yielded substantially different estimates of the relative risk reduction in bleeding associated with dabigatran and conflicting findings with regards to stroke. In subsequent analyses re-calculating the IPTW results, a hazard ratio of 0.57 (0.46–0.71) for the risk of stroke and 0.75 (0.69–0.82) for the risk of major bleeding was obtained, which were closer (but not identical) estimates to the PSM results. Therefore, as these methods become more widely used in real-world studies comparing the effectiveness and safety of OACs, there is an important need to improve understanding of the fundamentals of PSM and IPTW in order to cast a critical eye over study findings within the field of AF research.

To this end, this review describes PSM and IPTW methods to address confounding by indication in real-world studies, appraises similarities and differences between these techniques and alongside gives illustrative examples from some case studies comparing the effectiveness and safety of OACs. A clearer conceptualization of PSM and IPTW methods will help to ensure the findings of comparative studies are correctly interpreted, communicated and implemented into AF-related treatment decisions in clinical practice.

Case studies

Four case studies involving the use of PSM, IPTW or a combination of both methods in real-world CER of OACs will be presented throughout this article to demonstrate how these methods have been applied in practical terms. These case studies were selected from a range of researchers, data sources and geographies (Table 1). The ARISTOPHANES study from USA [23] and the SAKURA AF Registry study from Japan [24] used PSM to compare the effectiveness and safety of apixaban, dabigatran, rivaroxaban and warfarin among nonvalvular AF patients [23]. Using the Danish nationwide databases, Larsen and colleagues, [25] instead used the IPTW method to investigate the comparative effectiveness and safety of the same four OACs. Seeger and colleagues, [22] used both PSM and IPTW techniques to compare warfarin with dabigatran in two commercial health insurance databases in the USA.

Table 1. Selected case studies of real-world comparative effectiveness research of oral anticoagulants using propensity score matching and/or inverse probability of treatment weighting.

Study (year)	Data source	OACs	Main study end points	Variables included in propensity score	Statistical analysis	Ref.
Lip (2018)	ARISTOPHANES, USA	Warfarin, dabigatran, rivaroxaban, apixaban	– Stroke/SE, including ischemic stroke, hemorrhagic stroke and SE – Major bleeding, including gastrointestinal bleeding and intracranial hemorrhage	Age, gender, US region, Charlson Comorbidity Index, bleeding history, congestive heart failure, diabetes mellitus, hypertension, renal disease, liver disease, myocardial infarction, dyspepsia/stomach discomfort, peripheral vascular disease, stroke/SE, TIA, anemia/coagulation defects, alcoholism, concomitant use of: ACE/ARBs, amiodarone, beta-blockers, H2-receptor antagonists, proton pump inhibitors, statins, antiplatelet, NSAIDS	PSM: 1:1 propensity score matching. Nearest neighbor matching method without replacement with a caliper of 0.01, was used to match the patients	[23]
Okumura (2018)	SAKURA AF Registry, Japan	Warfarin, dabigatran, rivaroxaban, apixaban	– Stroke (ischemic stroke, hemorrhagic stroke or TIA) or SE – Major bleeding	Age, sex, height, weight, paroxysmal AF, hypertension, diabetes mellitus, heart failure, vascular disease, stroke or TIA, major bleeding, history of ablation, antiplatelet agent use, nonsteroidal anti-inflammatory drug use, OAC therapy duration and CrCl	PSM: matching was achieved by a 1:1 nearest neighbor approach (without replacement) within a caliper of 0.05	[24]
Larsen (2016)	Danish nationwide databases, Denmark	Warfarin, dabigatran, rivaroxaban, apixaban	– Ischemic stroke; a composite of ischemic stroke or systemic embolism; death; and a composite of ischemic stroke, systemic embolism or death – Any bleeding, intracranial bleeding, major bleeding	Age, sex; ischemic stroke or systemic embolism or transient ischemic attack; vascular disease; hypertension; diabetes; cancer; recent prescription of aspirin, beta-blockers, nonsteroidal anti-inflammatory drugs, or statins; and CHA2DS2-VASc and HAS-BLED scores	IPTW: weights calculated using generalized boosted models, based on 10,000 regression trees	[25,45]
Seeger (2016)	Two commercial health insurance databases (MarketScan® / Clinformatics®), USA	Warfarin, dabigatran	– Hospitalization for stroke (hemorrhagic or ischemic) – Hospitalization for major bleeding, including intracranial and extracranial bleeding	Age, sex, census region, calendar time, coronary artery disease, systemic embolism, DVT, PE, hypertension, diabetes, hyperlipidemia, atherosclerosis, heart failure, stroke, intracranial bleeding, hemorrhagic stroke, ischemic stroke, previous TIA, CHADS2 score, CHA2DS2-VASc score, previous MI, HAS-BLED, peptic ulcer disease, upper GI bleed, lower/unspecified GI bleed, urogenital bleed, other bleeds, peripheral vascular disease or PVD surgery, prior liver disease, prior cancer, renal dysfunction, acute renal disease, chronic renal insufficiency, diabetic nephropathy, hypertensive nephropathy, miscellaneous renal insufficiency, aspirin, aspirin/dipyridamole, clopidogrel, prasugrel, ticagrelor, other antiplatelet agents, NSAIDs, heparin, low-molecular weight heparins, PGP inhibitors, ARB, ACE inhibitor, Beta blocker, calcium channel blocker, other hypertension drugs, antiarrhythmic drugs (other than amiodarone and dronedarone), statin, other lipid-lowering drugs, diabetes medications, antidepressants, antipsychotics, anticonvulsants, proton pump inhibitor, H2 receptor antagonist, other gastroprotective agents, vitamin K therapy, number of medications, number of hospitalizations, number of hospital days, number of office visits, number of cardiologist visits, number of neurologist visits, hospitalization in 30 days prior to treatment initiation, number of laboratory tests ordered, number of INR (prothrombin) tests ordered, number of lipid tests ordered, number of creatinine tests ordered, treating prescriber	PSM: A 1:1 PS match was performed using the nearest neighbor methodology with a maximum caliper of 0.05 and additional matching within the calendar quarter to improve balance with respect to prescribing behaviors that might change over time IPTW: weighting was performed using IPTW (1/PS for dabigatran initiators and 1/(1−PS) for warfarin initiators). The IPTW analysis was repeated after capping (truncating) the weights at a value of 4.0.	[22]

ACE: Angiotensin-converting enzyme; AF: Atrial fibrillation; ARB: Angiotensin II Receptor Blockers; DVT: Deep vein thrombosis; INR: International normalised ratio; IPTW: Inverse probability of treatment weighting; NSAID: Nonsteroidal anti-inflammatory drugs; OAC: Oral anticoagulant; PE: Pulmonary embolism; PSM: Propensity score matching; SE: Systemic embolism; TIA: Transient ischemic attack.

Propensity scores: the basis for PSM & IPTW

Propensity scores form the basis for both PSM and IPTW methods, however as later described, the differentiating step is in how propensity scores are then used to control for differences in the characteristics of patients receiving the treatments being compared.

As defined by Rosenbaum and Rubin, an estimated propensity score reflects the probability of treatment assignment conditional on a patient's measured baseline characteristics, [26] such as their age, gender, comorbidities and concurrent medications. In the real-world setting, and in direct contrast to an RCT, treatments are not prescribed at random and are instead likely influenced by the characteristics of the patients at the time a treatment decision was made. For example, the decision to prescribe an OAC to an AF patient is likely informed by their stroke and bleeding risks. A patient's pretreatment characteristics can therefore be used to predict, and provide a measure (from 0 to 1) of, how likely the patient is to receive a treatment. Through leveling out the distribution of scores across treatment groups, propensity scores can minimize confounding by indication bias and provide a fairer comparison of different treatments [27].

In practice, propensity scores are estimated using regression-based methods and are easily implemented with statistical software packages such as SAS, Stata and R [28–30]. Most often, when comparing two treatments, a logistic regression model is used in which treatment assignment (a binary dependent variable with value 1 if the patient receives the treatment of interest and value 0 if the patient receive the comparator treatment) is regressed upon patient baseline characteristics (independent variables) [27]. When selecting variables to be included in the propensity score estimation, variables that are both related to the exposure (i.e., treatment assignment) and to the outcome (i.e., the main study end points) are recommended to provide the most precise estimates [31]. Misspecification of the propensity score can introduce bias and lead to invalid inferences therefore careful variable selection is essential. Extensions of the propensity score, such as the doubly robust method, facilitate more systematic variable selection and provide greater protection against model misspecification, however, are not yet widely known or adopted among researchers [32,33]. Three of the four case studies of this review used logistic regression models to generate the propensity scores, [22–24] whereas the fourth study used generalized boosted models, an advanced technique based on machine learning methods, applicable when comparing more than two treatments simultaneously [23]. Common variables included in the propensity score estimations were age, gender, bleeding history, stroke history, comorbidities and concomitant meditations (Table 1).

Once the propensity scores are estimated, a range of different ways can be used to balance the distribution of the scores, and in turn confounding factors, across treatment groups being compared. This includes propensity score adjustment, matching, stratification and weighting [27]. Among them, PSM and IPTW are emerging as the most commonly used propensity score methods within the field of AF research. Hence, the focus of this review is to provide foundational understanding of PSM and IPTW and highlight fundamental differences in methodology and interpretation of these two increasingly implemented techniques. Further resources on adjustment and stratification, and alternative weighting procedures to IPTW, are available elsewhere [27,34,35].

The main conceptual difference between PSM and IPTW relates to notion of the treatment effect being estimated. Whereas IPTW estimates the average treatment effect (ATE), PSM estimates the average treatment effect for the treated (ATT). The ATE reflects the effect of the treatment in the scenario that every patient within the population was offered the treatment. In essence, this shifts the entire population from untreated to treated. This contrasts with the ATT, which reflects the effect of the treatment only among those who were ultimately treated [27]. The research question being addressed will guide the decision as to whether an estimation of the ATE or ATT is of greater relevance. Estimation of the ATE may be less appropriate when a large proportion of patients are not good candidates for one of the treatment options. An example is the comparison of low dose apixaban patients with all warfarin patients, as some warfarin patients do not meet apixaban low dose criteria, it would be unrealistic to estimate the treatment effect among all patients [36]. Grasping the conceptual differences between the ATE and ATT is crucial to the correct interpretation of results from the application of PSM and IPTW.

PSM: estimating the ATT

PSM works by matching patients receiving the treatment of interest with patients receiving the comparator treatment based on the estimated value of their propensity score. PSM provides an estimation of the ATT, because only patients who received the treatment of interest are matched with comparators for comparative analyses [27].

In practice, patients are matched with others who have a similar propensity score value, as an exact score is not always possible. A variety of different matching methods are available [37] with four main analytics decisions:

Table 2. An overview of available propensity score matching techniques to pair up patients receiving the treatment of interest with suitable comparator patients.

Matching techniques	Definition	Considerations
Matching patients one-to-one (1:1) or many-to-one (m:1):		
1:1 matching	Matching patients 1:1 means that only one patient in the comparator group is selected for each treated patient	– 1:1 matching is most commonly used – m:1 has the advantage of retaining more patients in the analysis, however, the quality of second or third choice matches may be lower which increases bias
m:1 matching	Matching patients m:1 means that multiple comparator patients are selected for each treated patient	
Matching patients with or without replacement:		
Matching without replacement	Matching without replacement means that once a patient from the comparator group has been matched, they cannot be selected as a comparator for another treated patient	– Matching without replacement is more commonly used – Matching with replacement could be beneficial if the number of comparator patients is small, but may mean some comparators are matched multiple times and others not at all
Matching with replacement	Matching with replacement allows for a patient in the comparator group to be matched and act as a comparator for more than one of the treated patients	
Matching patients using a greedy or optimal technique:		
Greedy nearest neighbor matching	Using a greedy matching technique means that treated patients are randomly selected one at a time to be matched to their nearest comparator from the pool of remaining comparators still available to be matched. Here the closest match is made irrespective of whether the selected comparator would have been a better match for another treated patient	– Greedy nearest neighbor matching is more commonly used – Greedy and optimal matching in general perform similarly well in creating balanced treatment and comparator groups – Optimal matching may be preferred when the sample of comparator patients is small
Optimal matching	With optimal matching, the goal is to pair up treated and comparator patients such that the best possible combination of nearest matches are made	
Matching patients with a caliper width:		
Without a caliper width	When propensity score matching is performed without a caliper width, this means that matches are made without any limit on the distance between the propensity score values in each pair of matches	A caliper width is recommended to match patients within a given threshold. Without a caliper threshold, treated patients could be matched to their nearest comparator even if the differences in propensity scores were very large
Applying a caliper width	The role of a caliper width is to set a maximum distance between the propensity score values in each pair of matches	

matching patients one-to-one (1:1) or many-to-one (m:1) [38]; matching patients with or without replacement [39]; matching patients using a greedy or optimal technique [39]; and whether to apply a caliper width (Table 2) [40].

First, matching patients 1:1 means that only one patient in the comparator group is selected for each treated patient, as opposed to selecting multiple comparator patients in m:1 matching. The rationale for selecting more than one comparator is to make more use of the available data by retaining more patients in the analysis, which in turn may also lead to increased precision. However, matching patients m:1 has also been shown to increase bias because second or third choice matches may be of lower quality [39]. Matching patients 1:1 is most commonly used in practice [38,41].

Second, matching without replacement means that once a patient from the comparator group has been matched, they cannot be selected as a comparator for another treated patient. Conversely, matching with replacement allows for a patient in the comparator group to be matched and act as a comparator for more than one of the treated patients. Matching patients with replacement serves to increase the quality of matching and could be useful if there are few comparator patients relative to number of patients receiving the treatment of interest. However, with this approach it is possible that a single comparator patient could be matched multiple times, whereas another potential comparator may not be matched at all. This could result in the treatment effect being estimated based on a very restricted subset of comparators. Furthermore, when matching with replacement, the matched pairs are no longer independent and this must be accounted for in the subsequent analysis and interpretation of study findings [39]. Matching patients with replacements has seldom been implemented in practice [37].

Third, using a greedy matching technique means that treated patients are randomly selected one at a time to be matched to their nearest comparator from the pool of remaining comparators still available to be matched. Here the closest match is made irrespective of whether the selected comparator would have been a better match for another treated patient. In contrast, optimal matching intends to pair up treated and comparator patients such that the best possible combination of nearest matches are made. Optimal matching may be preferred over greedy matching when the sample of comparator patients is small, however when there is less competition for matches, optimal matching does not usually perform any better than greedy matching in creating balanced treatment and comparator groups. Greedy matching is more commonly used in practice [37,39].

Fourth, the role of a caliper width is to set a maximum distance between the propensity score values in each pair of matches. Without a caliper threshold, treated patients could be matched to their nearest comparator even if the differences in propensity scores were very large. In studies to date, there has been a lack of consistency in the maximum distance selected for the caliper width, however recent simulations support using a caliper width equal to 0.2 of the standard deviation of the logit of the propensity score [40].

Three out of the four case studies included this review implemented PSM techniques in their CER of OACs (Table 1) [22–24]. In the ARISTOPHANES study, Lip and colleagues used 1:1 nearest neighbor matching without replacement with a caliper width of 0.01 [23]. In the SAKURA AF Registry study, Okumura and colleagues also used 1:1 nearest neighbor matching without replacement, however, with a caliper width set at 0.05 of the standard deviation of the logit of the propensity score [24]. Similarly, in their analysis of two commercial health insurance databases in USA, Seeger and colleagues used 1:1 nearest neighbor matching with a caliper width of 0.05 and additional matching on calendar quarter to account for any changes in prescribing behaviors over time. It was not reported whether patients were matched with or without replacement [22]. The main difference between the three studies in terms of their chosen PSM approach is in the selection of the caliper width. The selection of a narrow caliper distance ensures patients are more closely matched on the value of the propensity score, however selecting a distance that is too narrow could result in too few matches being made [40].

Once the matched cohort of treated and comparator patients has been formed, the fundamental next step is to verify that a balanced distribution of patient characteristics across treated and comparator patients has been achieved by way of using the propensity score. This check could be made using statistical significance tests (e.g. X-square/*t*-test), however, computing the standardized differences of each baseline variable is now a more common practice. Lip and colleagues, and Seeger and colleagues both reported computing standardized differences with a threshold of 10% to check whether adequate balance of covariates across treatment and control groups had been achieved through the PSM process [22,23].

Were any systematic differences to be found between the treatment and comparator cohorts, the propensity score model may require some modifications on the included variables and/or matching procedure. Alternatively, subsequent adjustments could be made in the main analysis to account for any remaining imbalances in patient characteristics. Variables related to both the outcome and treatment assignment have been shown to provide the best balance of patient characteristics across treatment groups [31]. However, a common misperception for the development of the propensity score is to aim for perfect prediction of treatment assignment and the inclusion of variables that are related only to treatment assignment should be avoided [42]. Researchers should set out *a priori* which approach will be used in the event that balance in patient characteristics is not achieved through PSM (i.e., whether the propensity score will be re-estimated or if post-PSM adjustments will be made). Once satisfied with the PSM step, the comparative assessment can be performed, for example, comparing the risk of the study main outcome with hazard ratios from Cox proportional hazards regression models or some other analytic method of choice. What is advantageous about PSM is that the method matches patients who truly received the treatment with those who truly received the comparator to allow a direct comparison of the outcomes of these two groups. The method is transparent, readily understood and easy to communicate. One frequently discussed drawback of the PSM method is that patients who could not be matched are as a result excluded from the analysis. Unmatched patients may be systematically different from matched patients, limiting the representativeness of the study population and generalizability of the overall study findings. Yet, an alternative viewpoint is that this form of study population restriction ensures overlap across treatment and comparator groups on the most important patient characteristics thereby removing any individuals who are extreme outliers and arguably should not be compared [43]. Another consideration is that PSM is limited to pair-wise comparisons of one treatment against another. If more than one treatment comparison is being made, for example comparing apixaban versus warfarin and apixaban

versus dabigatran, the matched pairs for each of the comparisons are likely to be different, so drawing inferences across these two comparisons may not be appropriate.

IPTW: estimating the ATE

In the IPTW method, weights are assigned to patients based on the inverse of their probability of receiving treatment, as estimated by the propensity score. IPTW results in a pseudo-population in which patients with a high probability of receiving treatment have a smaller weight and patients with a low probability of receiving treatment have a larger weight and thus the distribution of measured patient characteristics used to calculate the propensity score becomes independent of treatment assignment. IPTW provides an estimation of the ATE, because the study population is re-weighted to assess the effects of the treatment in the scenario that it was offered to all patients within the population [27,44].

When comparing two treatments, the weight for each patient is calculated by inverting the probability of receiving the treatment the patient did in fact receive. A logistic regression model is usually used to calculate the propensity of receiving a treatment of interest versus a comparator. For patients in the treatment group, the weight is calculated as the inverse of the propensity score, whereas for patients in the comparator group, the weight is calculated as the inverse of 1 minus the propensity score (i.e., the probability of not receiving the treatment). Thus, the weight for a given patient j with propensity score p_j is calculated as:

$$\text{weight} = \begin{cases} 1/p_j & \text{if the patient is a member of the treatment group} \\ 1/(1-p_j) & \text{if the patient is a member of the comparator group} \end{cases}$$

Once calculated, the weights determine the extent to which each patient contributes to the new pseudo-population. For example, for a patient in the treatment group with $p_j = 0.25$, the weight is $1/0.25 = 4$, which represents four units in the pseudo-population. For a patient in the comparator group with $p_j = 0.25$, the weight is $(1/(1-0.25) = 4/3)$, which represents $4/3$ units in the pseudo-population. After the pseudo-population has been created, the balance of patient characteristics should be compared across the different groups using standardized differences, as in the case for PSM. Following this, the outcomes between treatment groups are ready for comparative assessment.

Two of the case studies presented in this review implemented IPTW techniques in their CER of OACs (Table 1) [22,23]. In addition to performing PSM, Seeger and colleagues also performed IPTW analyses by re-weighting the study population of dabigatran and warfarin users using the formula given above ($1/p_j$ for dabigatran and $1/(1-p_j)$ for warfarin) [23]. Larsen and colleagues, [25] on the other hand, used generalized boosted models to re-weight the study population of apixaban, dabigatran, rivaroxaban and warfarin users. Generalized boosted models can be used to calculate weights in the case when there are three or more treatment groups being compared simultaneously [45], however, this method, and others that are based upon machine learning techniques, are beyond the scope of explanation in this introductory review [46]. A simpler alternative, when there are three or more treatment cohorts is to compute the propensity score using a multinomial logistic model with all treatment cohorts (cohort #1, cohort #2, cohort #3, cohort #4) included in the model, using one cohort as the reference (i.e., cohort #1). Each patient's weight is equal to the inverse of the probability of receiving the treatment. The weight for a given patient j with propensity score p_j is calculated as:

$$\text{weight} = 1/p_j$$

For example, for a patient with $p_j = 0.5$ in cohort #2, the weight is $1/0.5 = 2$, which represents two units in the pseudo-population. For a patient with $p_j = 0.25$ in cohort #3, the weight is $1/0.25 = 4$, which represents four units in the full pseudo-population.

A key benefit of IPTW is that all eligible patients can be analyzed. This can be particularly useful when the study population is too small to afford to lose any treated patients who could not be paired with a comparator through a matching process. From a conceptual standpoint, IPTW is somewhat more difficult to comprehend and

Table 3. A side-by-side comparison of propensity score matching and inverse probability of treatment weighting: assumptions, advantages and disadvantages.

Method	Assumptions	Pros	Cons
PSM	• No unmeasured confounding • Positivity: every subject must have nonzero probability to receive either treatment • Correct model specification for propensity score • Compare two cohorts in most cases	• Easier to understand and communicate the data	• Excludes unmatched subjects who may differ systematically from matched subjects • Limit to independent pairwise comparison
IPTW	• No unmeasured confounding • Positivity: every subject must have nonzero probability to receive either treatment • Correct model specification for propensity score • Compare two or more cohorts	• Keep all eligible subjects • Can include more than two comparisons	• Difficult to understand and communicate the data • Less intuitive • Extreme weight issue

IPTW: Inverse probability of treatment weighting; PSM: Propensity score matching.

communicate. The question of what the ATE would be in the entire population, if the treatment were offered to every member of the population, could be of interest to healthcare payers. However, oftentimes a treatment is not suitable for all members of the population. Recall the example of comparing low-dose apixaban patients with all warfarin patients, as some warfarin patients do not meet apixaban low dose criteria they are therefore not good candidates for low-dose apixaban [36]. One methodological consideration for IPTW is the issue of extreme weights. This can occur when a treated patient has an extremely low propensity score, and then a very large weight is created. Large weights can increase the variability of the estimated treatment effect, leading to potentially biased results. In order to address this, stabilized weights should be used, which trim or truncate weights to a defined threshold [47]. Seeger and colleagues encountered the issue of extreme weights, when implementing the IPTW method to compare outcomes among NVAF patients treated warfarin versus dabigatran in USA. A hazard ratio of 0.00 (0.00–0.56) for the risk of stroke and 0.08 (0.08–0.10) for the risk of major bleeding was obtained with IPTW, contrasting with PSM estimates of 0.77 (0.54–1.09) and 0.75 (0.65–0.87) respectively. The IPTW analysis was repeated after capping (truncating) the weights at a value of 4.0, resulting in IPTW estimates of 0.57 (0.46–0.71) for the risk of stroke and 0.75 (0.69–0.82) for the risk of major bleeding, which were closer (but not identical) to the PSM estimated results (0.77 [0.54–1.09] for stroke and 0.75 [0.65–0.87] for major bleeding, respectively). Lastly, whereas PSM is restricted for pair-wise comparisons of one treatment against another, IPTW is feasible to include more than two treatment groups for a comparison.

Summary

PSM and IPTW are increasingly used to address confounding by indication potentially arising in CER due to lack of randomization in treatment assignment. These methods strive to emulate the properties of an RCT in creating treatment and comparator groups with balanced distributions of patient characteristics. Both methods, while based upon the propensity score, have different interpretations, which may be more or less suitable under different scenarios. In the case where there are ample numbers of comparators available for matching to treated patients, PSM offers a more transparent method, which is readily understood and easy to communicate. Furthermore, PSM may also be more robust to misspecification of the propensity score than the IPTW method, where extreme weights can bias the estimation of the treatment effect. On the other hand, IPTW has its advantages in retaining all eligible patients in the analysis, which may be preferred if there are limitations in terms of sample size, as well as the ability to include more than two treatment comparisons simultaneously. A side-by-side comparison of the two methods is detailed in Table 3.

Conclusion

To complement RCT evidence, the comparative effectiveness and safety of the four DOACs (apixaban, dabigatran, edoxaban and rivaroxaban) versus one another and against warfarin continue to be extensively researched in real-world studies. With PSM and IPTW being increasingly used in CER of OACs, as methods to achieve balance between treatment and comparator groups, this review provides an important introduction with applied examples from the AF field to aid understanding of these methodologies. A stronger foundational basis should help researchers and end users of CER of OACs correctly interpret, communicate and implement findings into AF-related treatment decisions in clinical practice.

Future perspective

As CER based on real-world data gains greater prominence in healthcare decision-making, it is crucial that decision-makers are familiar with emerging methodologies used in CER to rigorously assess the strengths, limitations and quality of the evidence in order to make informed decisions.

In the field of AF research, the comparative effectiveness of available OACs continues to be extensively researched with PSM and IPTW being two increasingly applied statistical techniques. This review provides an introductory guide describing these methods side-by-side to ensure that evidence from future CER studies of OACs are critically assessed, carefully interpreted and appropriately acted upon.

Executive summary

- After decades of warfarin being the only oral anticoagulant (OAC) widely available for stroke prevention in atrial fibrillation (AF), four direct OACs (apixaban, dabigatran, edoxaban and rivaroxaban) were approved after demonstrating noninferior efficacy and safety versus warfarin in randomized controlled trials (RCTs).
- Comparative effectiveness research (CER) of OACs based upon real-world data provides complementary information to the evidence provided by RCTs.
- In real-world studies, treatments are not prescribed at random, therefore confounding by indication bias may arise in CER if there are inherent differences in the patients prescribed the two or more treatments being compared.
- Propensity score matching (PSM) and inverse probability of treatment weighting (IPTW) are increasingly popular methods used to address confounding by indication in real-world CER of OACs.
- This review was undertaken to describe the fundamentals of PSM and IPTW and presents applied examples to assist researchers and end users of CER within the atrial fibrillation field of research in critically appraising and interpreting study findings.
- Key methodological considerations -
  - PSM and IPTW methods strive to emulate the properties of an RCT in creating treatment and comparator groups with balanced distributions of patient characteristics.
  - Propensity scores form the basis for both PSM and IPTW methods, however what differs is how propensity scores are used to control for differences in characteristics of patients receiving the treatments being compared.
  - PSM works by matching patients receiving the treatment of interest with patients receiving the comparator treatment based on the estimated value of their propensity score, creating pairs of treatment and comparator patients with a similar probability of receiving treatment.
  - PSM estimates the average treatment effect for the treated, reflecting the effect of the treatment only among those who were ultimately treated.
  - In the IPTW method, weights are assigned to patients based on the inverse of their probability of receiving treatment, as estimated by the propensity score, creating a new pseudo-study population where treatment assignment is independent.
  - IPTW estimates the average treatment effect, reflecting the effect of the treatment in the scenario that every patient within the population was offered the treatment.
  - Each method has its own practical advantages as well as limitations, which may be more or less suitable under different scenarios such as the availability of treatment and comparator patients and number of treatments being compared.
  - PSM and IPTW may not always lead to the same study conclusions therefore results should be viewed with full consideration of the technical differences in methodology and differences in the measurement and interpretation of the treatment effect that each method provides.

Author contributions

All authors substantially contributed to the development and critical revision of the intellectual content and approved the final version.

Financial & competing interests disclosure

This study was funded by Bristol-Myers Squibb and Pfizer Inc. All authors are employees of the funders. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

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

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Chugh SS, Havmoeller R, Narayanan K *et al.* Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study. *Circulation* 129(8), 837–847 (2014).
2. Miyasaka Y, Barnes ME, Gersh BJ *et al.* Secular trends in incidence of atrial fibrillation in Olmsted County, Minnesota, 1980 to 2000, and implications on the projections for future prevalence. *Circulation* 114(2), 119–125 (2006).
3. Ganjehei L, Massumi A, Razavi M, Rasekh A. Stroke prevention in nonvalvular atrial fibrillation. *Texas Heart Inst. J.* 38(4), 350–352 (2011).
4. Hart RG, Pearce LA, Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Ann. Intern. Med.* 146(12), 857–867 (2007).
5. Granger CB, Alexander JH, Mcmurray JJ *et al.* Apixaban versus warfarin in patients with atrial fibrillation. *N. Engl. J. Med.* 365(11), 981–992 (2011).
6. Connolly SJ, Ezekowitz MD, Yusuf S *et al.* Dabigatran versus warfarin in patients with atrial fibrillation. *N. Engl. J. Med.* 361(12), 1139–1151 (2009).
7. Giugliano RP, Ruff CT, Braunwald E *et al.* Edoxaban versus warfarin in patients with atrial fibrillation. *N. Engl. J. Med.* 369(22), 2093–2104 (2013).
8. Patel MR, Mahaffey KW, Garg J *et al.* Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N. Engl. J. Med.* 365(10), 883–891 (2011).
9. Kendall JM. Designing a research project: randomised controlled trials and their principles. *Emerg. Med. J.* 20(2), 164–168 (2003).
10. Berger ML, Sox H, Willke RJ *et al.* Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiol. Drug Saf.* 26(9), 1033–1039 (2017).
11. Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. *J. Multidisciplin. Healthcare* 11, 295–304 (2018).
12. Li G, Lip GYH, Holbrook A *et al.* Direct comparative effectiveness and safety between non-vitamin K antagonist oral anticoagulants for stroke prevention in nonvalvular atrial fibrillation: a systematic review and meta-analysis of observational studies. *Eur. J. Epidemiol.* 34(2), 173–190 (2019).

• **Meta-analysis of existing comparative effectiveness research on oral anticoagulants.**

13. Sterne JA, Hernan MA, Reeves BC *et al.* ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 355, i4919 (2016).
14. Joseph KS, Mehrabadi A, Lisonkova S. Confounding by indication and related concepts. *Curr. Epidemiol. Rep.* 1(1), 1–8 (2014).

• **Introductory reading on the concept of 'confounding by indication'.**

15. Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. *JAMA* 316(17), 1818–1819 (2016).
16. Schaumberg DA, Mcdonald L, Shah S, Stokes M, Nordstrom BL, Ramagopalan SV. Evaluation of comparative effectiveness research: a practical tool. *J. Comp. Eff. Res.* 7(5), 503–515 (2018).
17 . Borah BJ, Moriarty JP, Crown WH, Doshi JA. Applications of propensity score methods in observational comparative effectiveness and safety research: where have we come and where should we go? *J. Comp. Eff. Res.* 3(1), 63–78 (2014).

• **Commentary on the increasing application of propensity score-based methods in comparative effectivenss research.**

18. Biondi-Zoccai G, Romagnoli E, Agostoni P *et al.* Are propensity scores really superior to standard multivariable analysis? *Contemp. Clin. Trials* 32(5), 731–740 (2011).
19. Elze MC, Gregson J, Baber U *et al.* Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *J. Am. Coll. Cardiol.* 69(3), 345–357 (2017).
20. Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin. Pharmacol. Toxicol.* 98(3), 253–259 (2006).
21. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am. J. Epidemiol.* 158(3), 280–287 (2003).
22. Seeger JD, Bykov K, Bartels DB, Huybrechts K, Schneeweiss S. Propensity score weighting compared to matching in a study of dabigatran and warfarin. *Drug Saf.* 40(2), 169–181 (2017).

23. Lip GYH, Keshishian A, Li X *et al.* Effectiveness and safety of oral anticoagulants among nonvalvular atrial fibrillation patients. *Stroke* 49(12), 2933–2944 (2018).
24. Okumura Y, Yokoyama K, Matsumoto N *et al.* Three-year clinical outcomes associated with warfarin vs. direct oral anticoagulant use among Japanese patients with atrial fibrillation-findings from the SAKURA AF registry. *Circulation* 82(10), 2500–2509 (2018).
25. Larsen TB, Skjoth F, Nielsen PB, Kjaeldgaard JN, Lip GY. Comparative effectiveness and safety of non-vitamin K antagonist oral anticoagulants and warfarin in patients with atrial fibrillation: propensity weighted nationwide cohort study. *BMJ* 353, i3189 (2016).
26. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55 (1983).

•• **Propensity score original citation.**

27. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav. Res.* 46(3), 399–424 (2011).

•• **Introductory reading on propensity score matching and inverse probability of treatment weighting concepts.**

28. Gant T, Crowland K. A practical guide to getting started with propensity scores. https://support.sas.com/resources/papers/proceedings17/0689-2017.pdf
29. Leslie S, Thiebaud P. Using propensity scores to adjust for treatment selection bias. https://support.sas.com/resources/papers/proceedings/proceedings/forum2007/184-2007.pdf
30. Grotta A, Bellocco R. A review of propensity score: principles, methods, and application in Stata, Italian Stata Users' Group Meetings 2014, Stata Users Group. https://www.stata.com/meeting/italy14/abstracts/materials/it14_grotta.pdf
31. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *J. Thoracic Cardiovasc. Surg.* 134(5), 1128–1135 (2007).
32. Funk MJ, Westreich D, Wiesen C, Sturmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am. J. Epidemiol.* 173(7), 761–767 (2011).
33. Li X, Shen C. Doubly robust estimation of causal effect: upping the odds of getting the right answers. *Circ. Cardiovasc. Qual. Outcomes* 13(1), e006065 (2020).
34. Deb S, Austin PC, Tu JV *et al.* A review of propensity-score methods and their use in cardiovascular research. *Can. J. Cardiol.* 32(2), 259–265 (2016).
35. Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ* 367, l5657 (2019).
36. Nielsen PB, Skjoth F, Sogaard M, Kjaeldgaard JN, Lip GY, Larsen TB. Effectiveness and safety of reduced dose non-vitamin K antagonist oral anticoagulants and warfarin in patients with atrial fibrillation: propensity weighted nationwide cohort study. *BMJ* 356, j510 (2017).
37. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat. Med.* 33(6), 1057–1069 (2014).
38. Austin PC. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *Am. J. Epidemiol.* 172(9), 1092–1097 (2010).
39. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* 25(1), 1–21 (2010).
40. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceut. Stat.* 10(2), 150–161 (2011).
41. Rassen JA, Shelat AA, Myers J, Glynn RJ, Rothman KJ, Schneeweiss S. One-to-many propensity score matching in cohort studies. *Pharmacoepidemiol. Drug Saf.* 21(Suppl. 2), 69–80 (2012).
42. Bergstra SA, Sepriano A, Ramiro S, Landewe R. Three handy tips and a practical guide to improve your propensity score models. *RMD Open* 5(1), e000953 (2019).
43. Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report–Part III. *Value Health* 12(8), 1062–1073 (2009).
44. Austin PC, Stuart EA. 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(28), 3661–3679 (2015).
45. Mccaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* 32(19), 3388–3414 (2013).
46. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am. J. Epidemiol.* 185(1), 65–73 (2017).
47. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health* 13(2), 273–277 (2010).