



Advancing Analytical Methodologies for Unmeasured Confounding in Real-World Evidence

PHARMA-BIO WORKSHOP SUMMARY
May 2024



TABLE OF CONTENTS

1	Section 1: Introduction
5	Section 2: Defining Unmeasured Confounders
7	Section 3: The Potential of RWE
10	Section 4: Themes From the Workshop
14	Section 5: A Framework to Categorize Biostatistical Methods That Address Unmeasured Confounding
20	Section 6: Conclusion and Next Steps
22	References

SECTION 1: INTRODUCTION

The opportunity to use large amounts of real-world data (RWD) to evaluate the benefits and risks of therapeutic interventions has long been recognized by stakeholders across the health care ecosystem including regulators, academics, providers and the biopharmaceutical industry. As defined by the U.S. Food and Drug Administration (FDA), RWD are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.¹ Examples of RWD include data derived from electronic health records, medical claims data, data from product or disease registries and data gathered from other sources (such as digital health technologies) that can inform on health status. Real-world evidence (RWE) is clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD that can address important questions across the drug development lifecycle (from early research and development (R&D) through post-approval). Recent efforts to link data across sources to create robust population-level datasets and advancements in the creation and use of RWE are bringing us closer to realizing its full potential for evidence generation. These developments inspired nearly 500 stakeholders across industry, academia, regulators and patient advocacy to meet at a PhRMA-BIO workshop in Washington, D.C., on October 24-25, 2022, (the Workshop) to discuss one of the key challenges to the validity of RWE studies, uncontrolled confounding and innovative approaches to addressing it. The goal of the Workshop was to advance state-of-the-art analytical methodologies to address unmeasured confounders in RWE and to bring together professionals from diverse backgrounds to foster appropriate collaborations across stakeholder groups. This paper summarizes key insights and learnings on the methodologies from the Workshop and is not necessarily a reflection of PhRMA or BIO views.

RWE can be a useful tool across the drug development lifecycle, including discovery and early R&D, clinical development, pre- and post-approval regulatory requirements and supporting expanded uses. Significant advances in pharmacoepidemiologic research methods coupled with the increasing availability of rich, longitudinal health care data offer important opportunities to generate RWE that can inform health care decision-making. This evidence modality can be particularly helpful in situations where randomized controlled trials (RCTs) are not feasible or are unethical. In such situations, RWE may be generated in a more cost-effective and time-efficient manner. Among the many different use cases for RWE, the potential to inform on the benefit-risk profile of a medicine, either as part of a new drug application or in the post-approval setting (e.g., label expansion; safety), represents important opportunities. RWE can help make effective and safe treatments available to patients who currently have limited or no treatment options, including by serving as the basis for the FDA's determination of "substantial evidence" of effectiveness. RWE may also be an additional data source for safety and effectiveness information for approved therapies being administered post-marketing in real-world settings. In some cases, RWD may enable external validation of treatment effects in RCTs.

The use of RWE is not without challenges, many of which arise due to limitations of non-randomized studies that rely on health care data captured for purposes other than for research. The key attribute of RCTs is that bias is mitigated during the randomization of intervention and control groups to ensure that any imbalance of confounding factors, measured or unmeasured, is due to chance alone. In non-interventional RWE studies that leverage clinical practice data, treatment assignment is made by physicians based on patient characteristics (e.g., disease severity, formulary status, etc.), which introduces the potential for confounding.

Randomization is generally very effective at balancing patient characteristics and prognostic factors between treatment groups. Baseline exchangeability is expected, and consequently, the conventional frequentist statistics can be interpreted faithfully. Because of this, differences between groups can be causally attributed to the intervention itself and not to any other confounding variables. Confounding as a competing alternative explanation is a main limitation of RWE studies.

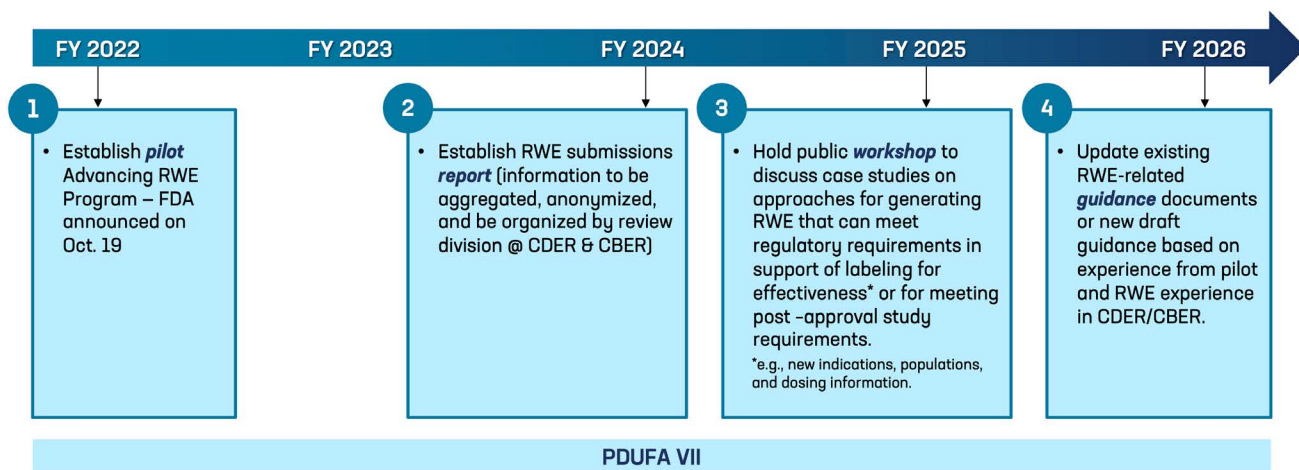
The inability to completely address or exclude the impact of confounding limits the trust and confidence that researchers and regulators can put into observational studies. Confounding is therefore a key challenge to the validity of non-interventional RWE studies examining the benefits and risks of therapeutic interventions and needs to be addressed in a rigorous manner in order to build confidence in RWE use.

The FDA recognizes both the promise of RWE as well as the associated challenges. As such, the 21st Century Cures Act and PDUFA VI created frameworks for the use of RWE and resulted in the issuance of guidance documents related to the use of RWE for regulatory activities.^{2,3} Building on these programs, RWE continues to be a priority for PDUFA VII. There are four key RWE initiatives under PDUFA VII through fiscal

year 2026 (see image below).⁴ Of these four, initiatives three and four – the FDA public stakeholder workshop to discuss RWE case studies and the FDA issuance of updated RWE-related guidance – are closely aligned with the objectives of the Workshop.

The biopharmaceutical industry is supportive of these initiatives to enhance the utility of RWE. Researchers within industry and academia, as well as regulators, share an interest in innovating and collaborating on novel methodologies within the established frameworks and in accordance with regulatory guidance. Thought leaders in epidemiology and biostatistics have worked to put forward novel methodologies to ameliorate the impact of unmeasured confounders on the interpretation of RWD/E. As these methodologies have improved and progressed, there has been a need for a public forum to discuss novel methodologies. Advances have been developed across a wide array of fields, including health outcomes research, biostatistics, epidemiology and economics. Often, researchers in one field may not be aware of advances in methods from other fields. This may lead to missed opportunities for the application of RWE, a lack of synergy across disciplines and less efficient approaches. The Workshop was intended to address some of these concerns.

RWE Under PDUFA VII



SECTION 2: DEFINING UNMEASURED CONFOUNDERS

RCTs have traditionally been the default method to determine causal treatment effects. Because of randomization, the risk of systematic bias is ostensibly removed at the time the initial population is assigned to the treatment and control groups. Additional bias may still creep in as a result of treatment progression, for example, resulting in different missing data patterns across treatment arms. In the real world, there may always be unmeasured confounders that introduce bias into the population that receive or do not receive treatment. This bias raises questions about the validity of an observational study's ability to attribute the patient's outcome to the treatment alone and not another variable. For example, a prescriber seeking the best outcomes for a patient population may select different drugs for a healthier patient cohort, assuming that these patients can better tolerate the side effects of the medication. Upon reviewing the outcomes of the population, a researcher is left with unclear attribution of the outcome to either the potency of the drug or the baseline characteristics of the cohort.

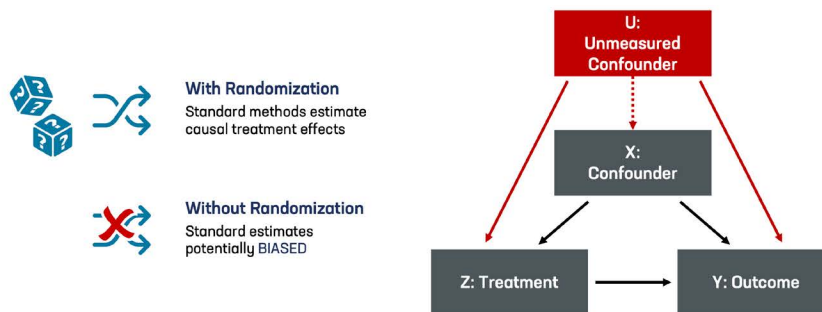
While the focus of the Workshop and this paper is on unmeasured confounding, it should be noted that it is one among other biases that can compromise the validity of research assessing the effect of a treatment using RWD. Other types of bias such as selection bias, immortal time bias or information bias can also bias the estimated treatment effect when using RWD.

Causal inference is of critical importance in drug development and regulatory evaluation.

Randomization is an ideal method to eliminate bias due to confounding. Historically, this has also been the only way to derive causal inferences. However, in RWD, because of a lack of randomization, confounding variables, especially unmeasured ones, can bias the treatment effect.

This potential for uncontrolled confounding can be illustrated through causal diagrams. At the outset of the Workshop, Dr. Doug Faries from Eli Lilly and Company introduced the classic causal diagram illustrating how confounders can impact both the treatment decision as well as the outcome (see image below).

In this diagram, one can see that a confounder, either measured (X) or unmeasured (U), can influence both the treatment decision (Z) as well as the outcome (Y). In practice, known and measured confounders are often explicitly controlled for. However, if the confounders are not accurately measured, adequately controlled for or unknown in the first place, researchers are left wondering what proportion of the outcome is attributable to the treatment and what proportion is attributable to the confounding variable. Note that the directed acyclic graph (DAG) above focuses on the challenge of bias through the non-random selection of treatment for a single exposure study. This diagram does not address potential biases that may arise in a longitudinal setting from treatment effects. These intercurrent events (ICEs) like dropout, medication switching or other confounding factors also have the potential for introducing bias.



Dr. Timothy Lash from Emory University used his meeting presentation to make a connection between quantitative bias analysis and confounding. He proposed characterizing unmeasured confounding as:

- **Unmeasured confounding:** The confounding variable is known, but cannot adequately be controlled because it is unmeasured.
- **Unknown confounding:** The confounding variable is not known and therefore no efforts to control the variable can be made by conventional approaches such as stratification, matching or regression.
- **Residual confounding:** The confounding variable is known and measured, but imperfectly measured or specified.



Since unmeasured confounding was the focus of the Workshop, below is a simplified example to further illustrate the concept.

- **Unmeasured confounding:** patient access to care
 - Access to care facilities can correlate with socioeconomic status. The research participant's home location may be potentially knowable, but not captured. The geographic distance to the nearest care facility could be derived if the home address was recorded. The travel distance is often further for those with lower socioeconomic status. To disentangle these two effects (travel distance being the confounder), one would need to be able to calculate the distance to available care facilities for each participant and determine the closest facility for every subject. Only then would it become clear whether principally socioeconomic status or distance to the nearest care facility drives the "true" differences between groups. Alternatively, it's possible that the patient's home address was originally captured ("measured") but travel distance was not derived because this type of confounding was unknown.

With the understanding that these types of unmeasured confounding can limit the conclusions drawn from RWE, investigators are developing various methodologies to overcome the limitations associated with unmeasured confounding present in observational research.

SECTION 3: THE POTENTIAL OF RWE

The success of advancing analytical methodologies to increase the utility of RWE will have a direct and positive patient impact. At the Workshop, Dr. Jeff Allen from Friends of Cancer Research and Annie Kennedy from the EveryLife Foundation represented the patient perspectives on RWE, discussing the critical need for robust RWE studies to illustrate the benefits and risks of new therapies. In oncology, only 5% of patients participate in clinical trials; by using RWD, researchers can assess data from a broader patient population. RWE can be used to help bring a personalized medicine approach to each individual patient or a protocol for groups of patients, allowing physicians to make more confident decisions on the best treatment for every patient on a case-by-case basis. For rare diseases, where patient populations are small, or for diseases with substantial unmet medical needs, where ethics limit options for placebo controls, strategies like external controls can reduce the size of clinical trials, potentially making infeasible trials feasible, and give more patients the opportunity to be randomized to the experimental treatment arm of a trial or give more patients the opportunity to receive experimental treatment in single-arm trials.

Expanding the scope beyond oncology and rare diseases, patients with other diseases can benefit from personalized medicine, or the ability to align

specific treatments to specific patients based on their biomarkers or other individual characteristics. Analysis of population databases can yield ever more refined patient cohorts to help prescribers make more informed decisions for individual patients. Prescribers rely on RWE to accurately assess expected benefits and risks between Drug A and Drug B, given an individual's personal health status. Increased accuracy for personalized medicine models benefits from large volumes of data that can be found in RWD, at levels that would never be economically feasible in RCTs alone. Personalized medicine can benefit patients across all clinical areas of medicine.

When appropriately applied, RWE can also help advance comparative effectiveness and safety research, optimize health care decision-making and improve patient outcomes. At the Workshop, Dr. Tzu-Chieh (Jay) Lin from Amgen highlighted key drivers in the growing demand for comparative effectiveness and safety studies using RWD, which include increasing acceptance by key stakeholders (i.e., regulators, payers and providers) as well as advancing data and analytic capabilities. Dr. Lin also emphasized that comparative effectiveness and safety research is challenging and requires a multidisciplinary team that can take a principled approach to designing and executing these studies.

Increasing Acceptance by Healthcare Authorities

- Regulators
 - External controls to support single arm filing strategies (oncology, rare diseases)
 - Comparative effectiveness of different treatments to support label expansion
 - Comparative studies to assess product safety in the post-marketing setting
- Payers and Providers
 - Comparative effectiveness and safety data to help guide [formulary] decision-making
 - Seeking data on benefits/risks of medicines outside of the "artificial" RTC setting

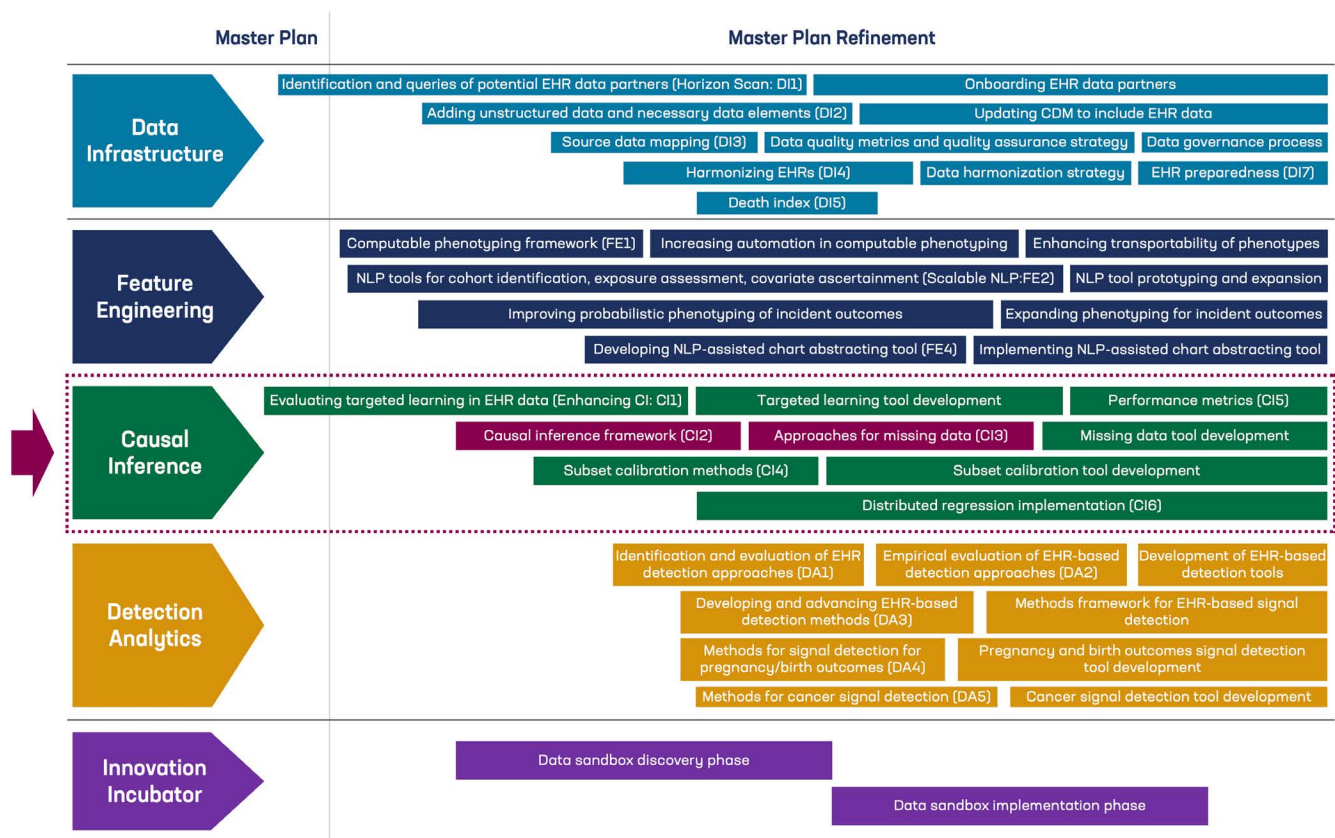
Data and Analytic Capabilities Are Advancing

- Regulatory RWE
 - FDA Sentinel System
 - DARWIN EU
 - ICH Working Group
- Access to large healthcare databases continues to increase (Optum Market Clarity [n~100M], Japan all-country claims, etc.)
- Analytic platforms that can execute CER studies (e.g., high-dimensional propensity score) are now available
- New standards are being broadly adopted

Indeed, increasing acceptance of such evidence has led to regulatory approvals, facilitating access to new treatments for patients. At the Workshop, Dr. John Concato from the FDA Center for Drug Evaluation and Research (CDER) highlighted a number of RWE-facilitated approvals over the years including:

- Approval of Blincyto in 2014 for the treatment of acute lymphoblastic leukemia based on data from a single-arm trial compared to patient-level historical data from chart reviews of patients at U.S. and EU sites.
- Approval of Zostavax in 2018 for the prevention of herpes zoster (shingles) in persons 50 years of age and older based on prospective, observational cohort study using electronic health records at Kaiser Permanente Northern California.
- Approval of Zolgensma in 2019 in patients less than two years of age with spinal muscular atrophy and a specific mutation based on data from a single-arm trial compared to data in an external control group from a natural history study.

FDA's Sentinel Innovation Center Portfolio



Beyond these three high-impact examples, additional examples were cited at the Workshop where RWE was used to inform regulatory decision-making and/or resulted in new drug approvals.

While these successes are important, regulators and other stakeholders continue to recognize the challenges associated with advancing the use of RWE to perform analyses and inform decision-making. Dr. Hana Lee from the FDA CDER summarized current agency initiatives intended to advance the use of RWE. In this plan, the initiatives under “Causal Inference” are the focus of the Workshop and this paper.

Dr. Lee helped to put into context the purpose of the Workshop with the broader FDA efforts across five domains to improve the utility of RWE and went on to highlight three key initiatives relevant to causal inference.



- 1 CDER public private partnership (PPP)**
 - RWE Scientific Working Group in the American Statistical Association (ASA) Biopharmaceutical Section: A PPP is established to address the issue of unmeasured confounding in generating and evaluating real-world evidence for regulatory purposes.
- 2 CDER Office of Medical Policy (OMP): Broad agency announcement RWE/D demonstration projects**
 - A Targeted Learning Framework for Causal Effect Estimation Using RWD
 - Detailing and Evaluating Tools to Expose Confounded Treatment Effects (DETECTe)
- 3 Method projects conducted under the Sentinel System**
 - The FDA is building Sentinel/ BEST methodology to improve understanding of robust evaluations with respect to RWE study design, analysis or variable measurement.

Building off Dr. Lee’s summary of the FDA’s efforts related to causal inference and unmeasured confounders, the Workshop then proceeded with a series of talks from leading industry and academic researchers on the latest biostatistical methodologies to advance the use of RWE.

SECTION 4: THEMES FROM THE WORKSHOP

Addressing unmeasured confounding is a top priority for building acceptance of RWE

There was broad alignment among all stakeholders at the Workshop that, when properly applied, high-quality RWE can inform better health care decisions. However, when it comes to specific use cases and analytical approaches in RWE, varying degrees of stakeholder hesitancy remain. Trust in RWE insights is paramount for the objective of broadening future adoption. The researcher community should continue to build and reinforce trust in RWE by “highlighting the good and critically evaluating the bad,” as Amgen’s Dr. Brian Bradbury summarized in his keynote address on October 24, 2022. Acknowledgment of what questions non-interventional approaches can and cannot answer is needed in order to build confidence in the use of RWD/E to answer certain research questions. The threat of unmeasured confounding is ever-present when working with RWD. Rather than relegating it to the “Discussion” section in academic papers where limitations of studies are discussed, unmeasured confounding should be addressed head-on as a first-order threat to validity in all non-interventional RWE studies. Researchers need to embrace the importance of addressing confounding in the most robust manner. As a matter of practice, RWE studies need to critically and objectively evaluate their success in this respect – or lack thereof. As part of the Workshop, an overview was given of ways to pursue this objective, and a preliminary framework was presented to guide researchers on appropriate methods. Dr. Michele Jonsson-Funk of the University of North Carolina at Chapel Hill cited in her talk that of 83 head-to-head non-interventional studies published between 2017 and 2019, only a third conducted formal analysis of unmeasured confounders.⁵ There is a pressing need to approach unmeasured confounding analysis more systematically and consistently. Quantifying uncertainty from unmeasured confounders will help

strengthen confidence in conclusions on the causal effects of treatment interventions.

A more thorough exploration of unmeasured confounding and quantitative approaches to assess bias (e.g., by using sensitivity analysis) could help build trust with regulators and allow for better understanding of the robustness of RWD as a data source to inform regulatory decision-making. Currently, there is no established consensus on when RWE is sufficiently rigorous to inform FDA regulatory decision-making. Advancements in analytical methodologies that address unmeasured confounding could also help advance FDA guidance on acceptable use of RWE in drug development and risk assessment for existing products.

While statistical approaches and specific methodologies may differ, the Workshop established stakeholder alignment on the importance of addressing unmeasured confounding in any future discussion of RWE methodologies.

Tackling unmeasured confounders requires cross-stakeholder collaboration

The RWD/E community is diverse – encompassing regulators, academia, patient advocacy groups, data service providers, as well as the biopharmaceutical industry. Each stakeholder brings a unique perspective to the table to advance the use of RWE. Unmeasured confounding as a challenge in RWE can be addressed through a baseline of commonly agreed upon approaches that address inherent risks when drawing causal inferences from non-experimental data, informed by each stakeholder’s unique perspective.

The Workshop provided a timely forum for the community to discuss and compare the merits and drawbacks of a variety of analytical approaches, including numerous emerging methodologies to deal with unmeasured confounding.

The discussions led to a call for “principled approaches” to foster trust in RWE. Also referred to as “structured approaches,” principled approaches promote consistency, ensure transparency and uphold the rigor of RWE analysis despite lingering challenges from unmeasured confounders. These considerations need to be made when studies are designed, not just post hoc at the time of analysis. Principled approaches would also provide an easily accessible and intuitive tool kit for less statistically equipped researchers to perform robust sensitivity analyses. With a broader and more experienced user base, RWD can be leveraged in more applications. Lastly, the creation of principled approaches provides a convenient framework for regulators such as the FDA to establish regulatory approval guidelines based on consensus best practices from the field.

Elements of a “principled approach” for designing and executing RWE analyses

Robust study design: Properly executed comparative effectiveness research starts with a comprehensive feasibility assessment. A thorough analysis, possibly enabled by mapping plausible causal diagrams, helps to identify measurable and unmeasurable potential confounders. Labeling and naming unmeasurable confounders helps highlight potential problematic project issues. This serves at least two possible purposes: it will help the interpretation of E-values later on (discussed further in Section 5) and might also give cause for reconsideration whether a project is feasible given potential unmeasurable confounding. This type of assessment requires collaboration across disciplines including medical experts, pharmacoepidemiologists and statisticians.

The goal of a successful RWE study, therefore, would be to emulate the outcome of a randomized study had it been conducted. Elements of the target randomized study being emulated can be explicitly characterized to aid the design of observational studies. In a paper published in 2016, Hernan and Robins outlined a

structured approach to the design of observational studies to avoid common methodological pitfalls.⁶ Core elements to be considered during the study design stage include:

- Eligibility criteria
- Treatment comparator arm
- Assignment procedures
- Follow-up period
- Outcome parameters
- Causal contrasts of interest (intent-to-treat vs. per-protocol)
- Estimands

Each element of the design should be carefully evaluated to account for the limitations of non-randomized observational data. For example, because patients are not assigned to each treatment arm randomly in RWD, we need to adjust for baseline confounders via methods such as matching, stratification and standardization.

In the Workshop, speakers and attendees shared their learnings on study design considerations, with a focus on statistical methodologies to help bridge the gap between observational and randomized studies. Although individual study situations may differ, the Workshop emphasized the need for a systematic process in study design and shared many examples to illustrate current best practices.

Fit-for-purpose data: The FDA has already published multiple draft guidance on data source considerations in RWE studies.⁷ Notably, data relevancy, provenance and quality were among the top considerations when selecting a data source. Because observational data in electronic health records (EHRs) and claims databases are not collected for research purposes, researchers should carefully assess the quality and appropriateness of the data with the research question in mind. Multiple data sources can and almost always are pooled together in a single study to create a purposefully specified comparator cohort.

What makes RWD fit for purpose? RWD are becoming increasingly available, and this has encouraged policymakers at the FDA and stakeholders in drug development to establish a set of considerations to determine whether data is fit for regulatory purposes.

Four key considerations come into play:

- 1 What is the regulatory question being considered?
- 2 What is the clinical context within which RWE is being generated?
- 3 What RWD of appropriate relevancy and quality are available?
- 4 What trusted methods are being applied to turn RWD into actionable evidence?

The data needs to be meaningful, valid, and the data provenance (sometimes also referred to as lineage) needs to be clear. The data needs to be representative of the population of interest, with adequate coverage of end points and covariates. This evaluation is particularly important when RWE is going to be used to support causal claims. RWD, because it is typically collected to record patient health status and/or delivery of care, has the potential for selection bias. The data set needs to be representative of the population of interest, have all critical fields available pertaining to exposures and outcomes and as many covariates as possible.

Data quality pertains to accuracy, completeness, but also provenance. Often, fields need to be imputed. In such cases, insight needs to be given in methods used to derive and transform data. Completeness can be delicate especially when patterns in missing data reveal they are not missing at random (NMAR). Provenance is also important so that users can follow

the audit trail to upstream sources. The selection process for data sources plays a key role because this allows an evaluation of potential systematic bias – one of the primary challenges to validity.⁸

Among the current best practices shared by speakers during the Workshop, a unifying theme emerged for data selection: datasets should be evaluated specifically in the context of the research question with a standardized procedure and defined “go/no go” criteria. Workshop speakers further noted that industry teams should include a formalized review process by an internal governing body specialized in observational research to determine if the data is fit-for-purpose in order to address the research question. These speakers also observed that data providers should continue to provide high quality, regulatory and research-grade RWD to support evidence generation and that raw clinical and claims data typically require extensive aggregation, translation, curation and linking to fulfill the depth and breadth requirements for each study. Speakers highlighted the importance of close collaboration between various players of the data ecosystem – hospitals, data aggregators, researchers in academia and industry – as essential to safeguard the rigor of RWE studies under the guidance of the FDA on a set of mutually agreed-upon standards. Making this selection process more transparent and auditable will help increase trust in the use of RWE.

Prespecified analyses: Given the complexities of RWE and the multitude of potential confounding factors, statisticians often need to select appropriate methods among a very large number of possible options for each study. A recent paper by Dr. Xiang Zhang – one of the closing speakers at the Workshop – and colleagues noted that in observational studies using RWD, “a large number of associations could often be found, when in reality only a few are true associations.”⁹ False positive associations may well be the result of data dredging. Data dredging, also known as “p-hacking”, is the cherry-picking of promising findings leading to a false excess of statistically significant results.¹⁰

This could lead to the misconception that non-interventional studies can only be analyzed in a post-hoc manner and therefore cannot achieve the same statistical robustness as randomized controlled trials. Dr. Zhang concluded the Workshop by issuing a call to action to the researcher community to “prespecify sensitivity analysis for unmeasured confounding” and “carefully distinguish prespecified analysis from ad-hoc, data-driven analyses” when reporting results from observational studies.

Prespecification is an important statistical principle outlined in ICH E9 in a section on documenting estimands and sensitivity analysis in the trial protocol and analysis plans.¹¹ Biopharmaceutical companies, regulatory agencies and clinical research organizations have been improving prespecification to explicitly address intercurrent events. The complement for RWD also requires this type of attention to ensure comparability and completeness between research arms. With prespecification, RWD could be collected and/or analyzed prospectively in studies, as opposed to being relegated to solely being retrospective. Data dredging would also be avoided through prespecification. Further adoption of this principled approach will likely improve the validity of comparative efficacy and safety conclusions generated by RWD/E.

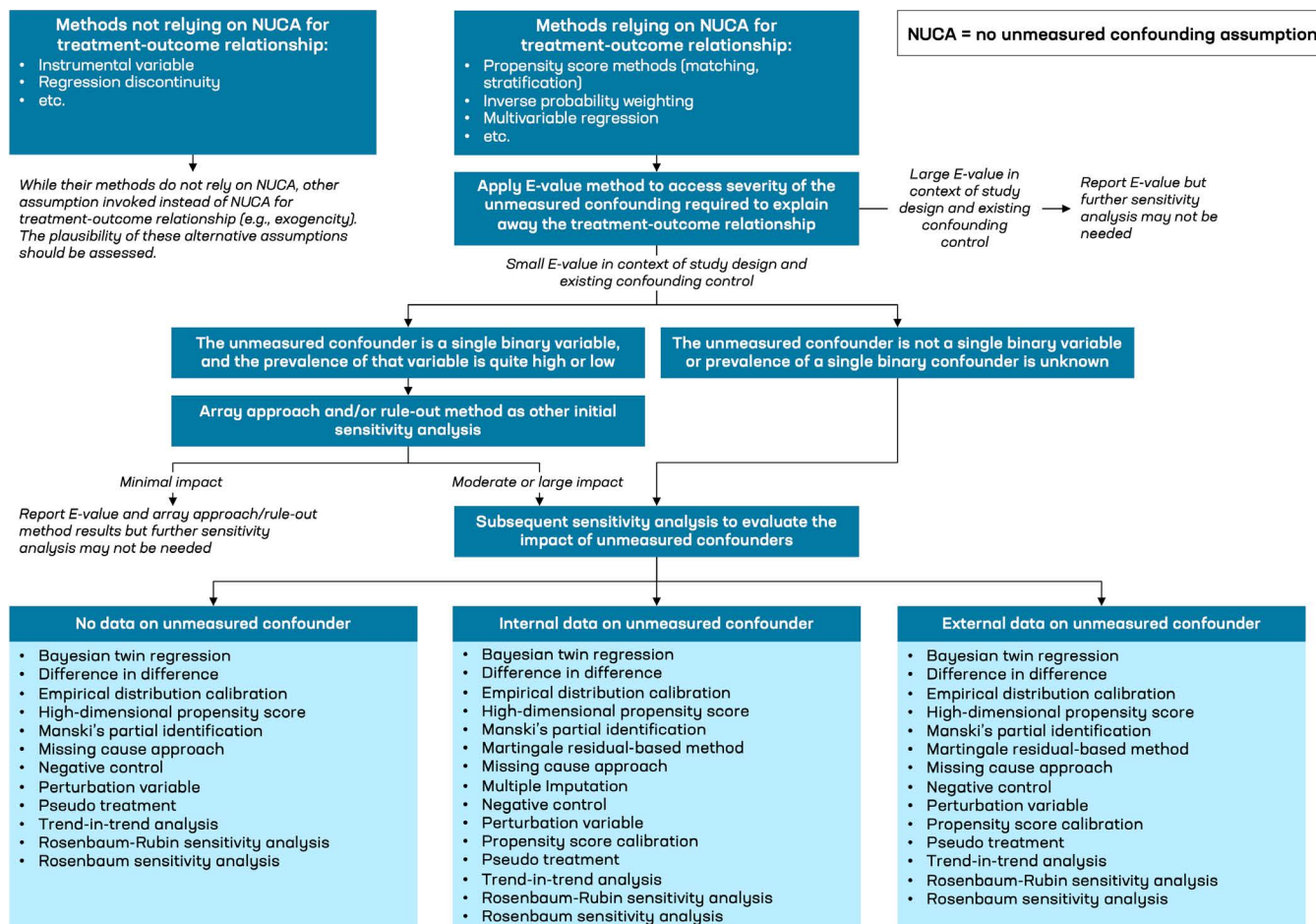
Unfortunately, RWD prespecification has not been widely adopted in all parts of the researcher community to date. One challenge to the prespecification of analyses is a lack of structured guidelines for selecting appropriate statistical methods to address unmeasured confounders. This was an important driver for convening the Workshop.

All methods in this category depend on assumptions of varying strengths to emulate a randomized trial and infer causality from observational data. The Workshop built upon preexisting efforts to propose a consensus framework for systematically evaluating and selecting appropriate statistical methods. We will provide a brief overview of the proposed framework in the next section of this paper while noting that its refinement and eventual adoption will require a collaborative effort from diverse perspectives in the RWE community.

Rigorous quality control: With a prespecified analysis plan in place, researchers need a method to determine whether current methods have sufficiently addressed unmeasured confounders. The magnitude of residual confounding reflects the upper bound of effects that are unaccounted for, and hence the unmeasured confounders may still have on the specified treatment effects. A rigorous framework is needed to assess the “quality” of an analytic approach and serve as a decision rule to determine whether a study is adequately robust in light of potentially unmeasured confounders. Different stakeholders may have widely different demands from any given study, in light of their respective needs. For example, patient advocacy groups may have different concerns than those of the FDA review staff. The Workshop featured a session on “Methods That Evaluate the Robustness of Study Findings,” where speakers shared methodologies, current best practices and example use cases. These diagnostic tools are capable of testing the validity of analysis plans and are essential components of a principled approach to deal with unmeasured confounding.

SECTION 5: A FRAMEWORK TO CATEGORIZE BIOSTATISTICAL METHODS THAT ADDRESS UNMEASURED CONFOUNDING

Suggested Steps to Evaluate the Impact of Unmeasured Confounders



Reproduced from Zhang et al. 2020

The Workshop organized presentations on various methodologies following a framework proposed by Dr. Zhang and colleagues, who summarized their recommendations in a flowchart depicted above. The flowchart is one of the earliest prototypes of a structured approach towards the selection of appropriate statistical methods to perform sensitivity analysis. Although not quite comprehensive, the framework marks an important step towards consensus-building in the RWE community. To guide the community towards a consensus approach, the authors invited feedback and discussion at the Workshop, as well as proposals for alternative frameworks. All these methods assume that the source data – especially accuracy in end points – are of sufficient quality to justify additional efforts towards data validation. During the Workshop Q&A session, participants suggested ways to improve evaluating the impact of unmeasured confounders.

The methodological sessions of the Workshop opened with quasi-experimental methods that do not rely on “no unmeasured confounding assumption” (NUCA) (see upper left corner of the flowchart). Unlike the majority of methods discussed at the Workshop, these methods circumvent the issue of unmeasured confounding without NUCA dependency. However, in the absence of true randomization, no method can gain advantages “for free.” This is analogous to Abelson’s “no free lunch” theorem.¹² Quasi-experimental methods rely on alternative assumptions, such as the availability of exogenous instrumental variables. Instrumental variables act as pseudo-randomization factors. They determine which treatment subjects receive, but have no causal connection to outcomes.

An important assumption is that instrumental variables also require independence from unmeasured confounding. Dr. Jian Cheng from University of California, San Francisco and Dr. Rishi Desai from Harvard University discussed instrumental variables at the Workshop through theory and by providing practical examples. Importantly, practitioners should carefully assess the plausibility of alternative assumptions required by these methods prior to deploying them. Specifically, instrumental variables need to induce substantial variation in the treatment variable but should not have any direct effect on the outcome variable of interest. They can be thought of as a device that achieves pseudo-randomization. The use of instrumental variables allows observational studies to serve as substitutes for randomized controlled trials (RCTs).¹³

Subsequent sessions of the Workshop explored sensitivity analysis methods for statistical models that do require NUCA. These methods could be broadly segmented into “initial” and “subsequent” based on the goals of analysis and the depth of assumptions required. Initial sensitivity analysis methods aim to test the robustness of study findings in light of potential unmeasured confounding with few additional assumptions. These “initial” methods provide a worst-case upper bound scenario to determine whether more comprehensive sensitivity analyses are required. In other words, if the strength of unmeasured confounding is lower than the bound, then the treatment effects are likely robust; no further analysis is needed. If the bound is exceeded, the validity of the observed treatment effect is threatened.

E-value is a tool researchers could use as part of initial sensitivity analysis to characterize robustness to residual confounding. Recently proposed by VanderWeele and Ding, the E-value captures the minimum strength of association that unmeasured confounders must have with exposure and outcome to fully explain away the observed effect.¹⁴ It requires minimal assumptions regarding the structure of unmeasured confounding and is straightforward to calculate with existing software packages. Dr. Maya Mathur of Stanford University provided an introduction to E-values during the Workshop, including examples and a list of frequently asked questions on their usage and interpretation. Because of its simplicity, the E-value approach can be effectively used as an initial triage for the severity of unmeasured confounding. A large E-value suggests that treatment effects are robust to unmeasured confounding, while a small E-value is an indicator that further analysis is needed. Because E-values do not make assumptions regarding the prevalence or distribution of unmeasured confounders, it is a conservative measure, and not all confounders with strengths surpassing the E-value are capable of explaining away the causal effects. Given these advantages, Dr. Mathur encouraged the researcher community to “routinely report E-value in observational studies” to “better calibrate confidence in causal effects.” As good practice, RWE studies are recommended to report this value, including the associated confidence interval for it.

Other speakers in the session provided alternative methodologies for initial sensitivity analysis while incorporating substantive knowledge of the unmeasured confounder. Dr. Desai from Harvard presented the array approach and rule-out methods, which can be used if the unmeasured confounder is a single binary variable, regardless of whether the prevalence of that variable is either quite high or low. In this example, socioeconomic status (SES) was appropriately ruled out as a potential confounder because the research findings showed that SES was more than two-fold higher in the reference group

and hence did not threaten the validity of the causal effect.¹⁵ Dr. Lash presented methods that incorporate prior knowledge into quantitative bias analysis to improve estimates beyond simple bounding. He introduced principles of bias quantification through design, such as incorporating the exposure prevalence in both experimental and control groups to correct for selection biases.

If initial sensitivity analysis using E-value, rule-out and/or array methods suggests that weak unmeasured confounding could meaningfully reduce the validity of the observed association, subsequent sensitivity analysis involving additional assumptions may be needed. The armamentarium of statistical methods for subsequent sensitivity analysis is large and evolving. No method is appropriate under all scenarios or for every study. Highlighting the need for additional guidance on selecting the right method for individual use cases, Zhang et al. proposed the following segmentation frameworks, which were reiterated by one of the coauthors and speaker at the Workshop, Dr. Faries.¹⁶

Segmentation by availability of information on the unmeasured confounders:

- 1 No information:** No additional information nor reasonable assumptions on the unmeasured confounder(s).
- 2 Internal information:** Information on the unmeasured confounder(s) is available from internal data (i.e., a subset of patients within the current study).
- 3 External information:** External data may contain information regarding the unmeasured confounder(s).

Segmentation by the goal of unmeasured confounding assessment:

- 1 Plausibility assessment:** The methods used to test whether the conclusion is insensitive over a range of plausible a priori assumptions on the unmeasured confounders. These “indirect methods” provide evidence on the influence of unmeasured confounders without directly providing guidance on adjusting the treatment effect estimate.
- 2 Adjusted sensitivity analysis:** The methods providing adjusted estimates of the treatment effect while controlling for the unmeasured confounding. These “direct adjustment methods” leverage internal or external data by invoking additional assumptions to directly calibrate causal effect estimates.

This segmentation provides valuable guidance on how to choose appropriate methods of sensitivity analysis. Researchers need to thoroughly evaluate the nature of the unmeasured confounders, the availability of extra information and the goals of the assessment before finalizing the analysis plan. Certain methods may only be appropriate when the confounders are measurable, while others can be implemented regardless of measurability. Furthermore, each methodology carries its own set of additional assumptions that need to be satisfied for appropriate use. There are certainly situations where multiple analytical methods are applicable, and researchers could apply more than one method as long as all analyses are appropriately documented and reported.

Speakers at the Workshop presented various analytical methods suitable for diverse settings and their associated applications. Dr. Arman Oganisian from Brown University spoke about Bayesian approaches to causal inference, both parametric as well as non-parametric optimization methods. He compared Frequentist with Bayesian model specifications and illustrated how bootstrap and Markov chain Monte Carlo are implemented in the latter. Dr. Thomas Jemielita from Merck presented a series of case studies where RWD was leveraged in hybrid randomized controlled trial settings to reduce bias compared to single-arm designs. Other companies have initiated similar hybrid Phase 3 development programs. Dr. Jessica Franklin from Optum shared her work on high-dimensional propensity scores (hdPS). Traditional approaches to covariate adjustment may be subjective as they depend on expert knowledge. In contrast, hdPS automates the identification, prioritization and adjustment for a large number of potential confounders. In a simulation study, hdPS consistently outperformed direct adjustment methods using regularization, and the base case of hdPS was so impressive that it was hard to improve upon further.¹⁷

Dr. David Lenis from Aetion explored the mechanisms of missing data and key considerations around handling missing data. Possible mechanisms are missing completely at random, missing at random and missing not at random. Depending on the mechanism at play, performing only complete case analysis or multiple imputations could lead to bias. He recommended that researchers should carefully consider the substantive mechanism that leads to missingness, as well as whether the information behind missing data is already accounted for in the observed data. When performing inference, researchers should explicitly state any inclusion criteria based on data completeness, document in detail methods used to handle missing data, and crucially, conduct sensitivity analysis to estimate the potential impact of missingness.

Dr. Mark Weiner from Weill Cornell Medicine spoke about the prior event rate ratio as a tool to tease out the treatment effect from baseline event rates. Although simple and effective, the method is relevant to specific types of medical interventions where the timing of treatment initiation is arbitrary and not triggered by an acute observation. In an example, the effect of statin treatment on cardiovascular events was studied using prior event rate ratios, since the initiation of statin treatment was not triggered by an immediate event.¹⁸ Dr. Jianchang Lin from Takeda shared a novel methodology that combines propensity score methods with meta-analytic priors. The method enables teams to simultaneously leverage external data adaptively as with meta-analytic predictive priors, while adjusting for patient-level covariates captured with propensity scores.^{19,20} Dr. Ting Ye from the University of Washington shared her work on instrumented difference-in-differences. Dr. Ye illustrated her method with the classic example of the causal connection between smoking and lung cancer. In the mid-20th century, women began smoking more as a result of advertising. Thirty-five years later, there was a noticeable increase in lung cancer mortality, specifically in women. Dr. Ye connected these two phenomena to illustrate how instrumental variables are used in this method.²¹ Dr. Satrajit Roychoudhury of Pfizer provided examples using meta-analytic predictive priors derived from external control data to augment the analysis of a single-arm Phase II trial.

The next session at the Workshop focused on “indirect methods” that assess the robustness of causal inference. These methods are examples that fulfill the “quality control” requirement from the overarching

principled approach proposed above. The session opened with presentations from Dr. Alan Brookhart from Duke University and Dr. Tzu-Chieh (Jay) Lin from Amgen. Dr. Brookhart introduced the concept of negative control outcomes, variables that are believed to be unaffected by the treatment but share the same confounding structure – both measured and unmeasured – with the outcome of interest. If a non-negligible association between the intervention and the negative control outcome is observed, then the analysis may be subject to residual confounding. Negative controls are the focus of an FDA workshop and related development projects in PDUFA VII, and an FDA report on their appropriate use and development is anticipated.²²

In the Q&A, numerous open questions were discussed, including how many negative control outcomes should be incorporated into a study when multiple are available, how to interpret the results from numerous negative control outcomes and whether to include multiplicity testing correction. Dr. Lin followed with a real-world comparative safety study where negative control outcomes were leveraged as a gating process after inverse probability of treatment weighting to quantify residual bias and inform the appropriate choice of active comparator.²³ In Dr. Lin’s study, no substantial bias was detected, and the conclusions of the study were therefore deemed to be robust.²⁴ Dr. Patrick Ryan from Janssen expanded upon the use cases of negative controls from bias quantification to estimand correction. Dr. Ryan and colleagues proposed empirical calibration methods to calibrate both p-values and confidence intervals.²⁵

Post-calibration, the expected absolute systematic error could be used as a diagnostic to determine whether residual confounding observed from negative controls is small enough to accept the calibrated effect estimates as unbiased. Dr. Bo Lu from Ohio State University concluded the session by clarifying the distinction between primal and simultaneous sensitivity analysis. The latter allows researchers to quantify how strong the effect of unmeasured confounders would need to be, in order to explain away the observational effects in a study. Dr. Lu's talk concluded the Workshop's section on statistical methods involved in subsequent sensitivity analysis and the principled approach framework.

The final session of the Workshop focused on one important application of RWD – the augmentation of clinical studies with external control arms, either in the form of single-arm studies with purely external controls or hybrid studies with a small concurrent control arm augmented by external subjects. Dr. John Seeger from Optum laid the groundwork for this session by introducing many types of unmeasured confounder threats to data validity, including selection bias and information bias, among other internal and external validity threats. Different data-generating mechanisms (e.g., claims or EHR versus clinical trials) could also contribute to bias from external comparison groups.

Dr. Mingyang Shan from Eli Lilly and Company expanded on this topic by sharing findings from a recent simulation study where different statistical methods were applied to analyze clinical trial data with hybrid control arms. Dr. Shan and colleagues concluded that although both Frequentist and Bayesian methods could perform well under violations of certain assumptions, no method was able to fully mitigate bias when unmeasured confounding variables are correlated with outcomes. These findings underscore the need to consider outcome adjustments when modeling and the importance of sensitivity analysis when borrowing from external controls. Dr. Laura Fernandes from COTA continued the discussion on the advantages and drawbacks of external control arms. External control arms could potentially be a solution to RCT's lack of external validity, though a valid estimate of treatment effect using RWD needs to be carefully calibrated in light of unmeasured confounding and biases. Dr. Michael Bretscher from Roche continued the discussion by presenting a meta-analytic framework to quantify and control for biases in external control studies.²⁶ He reiterated the need to prespecify an analysis plan for bias adjustment and that formulating standardized criteria for identifying fit-for-purpose historical reference studies remains an open challenge.



SECTION 6: CONCLUSION AND NEXT STEPS

The two-day Workshop brought together leading biostatisticians and epidemiologists to discuss both progress and challenges in addressing unmeasured confounding in non-interventional research. The event served as a forum to share innovative ideas and approaches, as well as to motivate continued innovation in this space with the ultimate goal of increasing the validity of non-interventional research. While the speakers discussed numerous statistical methods, the Workshop was not able to cover all emerging and available methods. The Workshop aimed to provide a starting point for further collaboration and alignment building within the researcher community and between various stakeholders across communities involved. A unified framework categorizing various methodologies and the principled approach presented at the Workshop and within this paper represent important first steps toward addressing that challenge. In addition, Workshop presenters and participants identified multiple next steps to further advance the use of RWE and adoption of RWE as a legitimate means to support and strengthen research findings, as described below.


Continue to foster broad stakeholder collaboration

A key feature of the Workshop was participation by various stakeholders across the R&D ecosystem, including regulators, industry, academics and patient advocacy groups, with a shared vision of advancing the use of RWE. Collaboration across these groups will be instrumental in advancing the field. The learnings from this Workshop were an important and valuable step forward to accelerate collaboration on and acceptance of RWE. The biopharmaceutical industry acknowledges the FDA's efforts to incorporate RWE into regulatory decision-making and looks forward to continued engagement with the agency on those initiatives, including those outlined in PDUFA VII.

Innovation from leading academic researchers can be applied to practical applications in drug development, driving a virtuous cycle that will advance the field. The FDA's recent initiative to conduct a public workshop on the use of negative controls²⁷ is an illustrative example of positive collaboration with stakeholder groups in the drug development ecosystem. Patient advocacy groups, among other stakeholders, continue to identify unmet medical needs and remind the field of the ultimate objective of these efforts, which is to improve patients' lives. Payers, who play a key role in access, utilization management and reimbursement of approved medicines, also have a perspective on the use of RWE. Involving and aligning all stakeholders in further efforts to advance the field is key to success.

Refine and align on a principled approach for RWE analysis and biostatistical methods to address unmeasured confounding

As the number of biostatistical options for RWE analytics continues to increase, ensuring appropriate use of these methodologies is critical and linked to ensuring trust in the conclusions of the analyses. A consensus principled approach for the selection and implementation of these methodologies will be valued by the RWE community and will increase confidence that various methodologies have been used appropriately. Indeed, many attendees of the Workshop expressed the need for a practical guide on when and how to use each method, which should increase the use of quantitative sensitivity analyses to assess the potential impact of unmeasured confounding. This paper presents a prototype of one version of such approach, though further refinement and alignment are needed. The Workshop participants hope that the field continues discussions of the circumstances, strengths and limitations behind each method in future fora.



A consensus practical guide on methods to address unmeasured confounding would help advance the use of RWE in improving drug development, regulatory decision-making and patient care.

Acknowledge residual uncertainty due to unmeasured confounding and incorporate sensitivity analyses in all RWE studies

Clinical evidence needs to meet a high threshold of validity to drive decision-making, as physicians rely upon the data to select treatments for patients. RWE has the potential to meet this threshold, provided researchers adhere to current best practices with respect to the initial study design and downstream analysis, which incorporates sufficient controls for dealing with confounding, including the potential for unmeasured confounding. This should help lead to reliable and valid outcomes that are necessary to strengthen trust in RWE and increase its appropriate use.

Unmeasured confounding remains the biggest threat to the validity of RWE, and the field must make addressing it a top priority. To this end, researchers conducting RWE studies should recognize the importance of addressing confounding and also accept that there are and always will be limitations and questions that cannot be answered using non-interventional approaches. Celebrating and highlighting positive use cases of RWE while critically evaluating the limitations of RWE will help to build trust in the outcomes from these studies.

The Workshop and this paper aim to improve awareness of methods to address unmeasured confounding. Statistical methods that depend on the No Unmeasured Confounding Assumption (NUCA) should not be accepted at face value without a thorough exploration of the potential impact of unmeasured confounders. Consistent with various best practice documents and modern epidemiology textbooks, ensuring the validity of non-interventional research through the robust assessment of confounding and the use of quantitative bias analysis will help improve use of RWD/E.

REFERENCES

- 1 Office of the Commissioner. (2019). Real-World Evidence. U.S. Food and Drug Administration. <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>
- 2 21st Century Cures Act. Energy and Commerce Committee. <https://energycommerce.house.gov/21st-century-cures>
- 3 Completed PDUFA VI Deliverables. U.S. Food And Drug Administration. <https://www.fda.gov/industry/prescription-drug-user-fee-amendments/completed-pdufa-vi-deliverables>
- 4 PDUFA VII: Fiscal Years 2023-2027. U.S. Food And Drug Administration. <https://www.fda.gov/industry/prescription-drug-user-fee-amendments/pdufa-vii-fiscal-years-2023-2027>
- 5 Latour, C., Poole, C., Edwards, J., Sturmer, T., Martin, D., Lund, J. L., ... & Funk, M. J. (2020). Use of sensitivity analyses to assess uncontrolled confounding in observational, head-to-head pharmacoepidemiologic studies: A systematic review. *Pharmacoepidemiology and Drug Safety*, Vol. 29, 382-382.
- 6 Hernán, M. A., & Robins, J. M. (2016). Using Big Data to Emulate a Target Trial When a Randomized Trial is Not Available. *American Journal of Epidemiology*, 183(8), 758-764. <https://doi.org/10.1093/aje/kww254>
- 7 FDA. (2023). Real-World Data: Assessing Support Regulatory Decision-Making for Drug and Biological Products. <https://www.fda.gov/media/154449/download> FDA. (2021). Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products, draft guidance. <https://www.fda.gov/media/152503/download>
- 8 Daniel, G., McClellan, M., Silcox, C., Romine, M., Bryan, J., Frank, K. (2018). Characterizing RWD Quality and Relevancy for Regulatory Purposes. Duke Margolis Center for Health Policy. https://healthpolicy.duke.edu/sites/default/files/2020-03/characterizing_rwd.pdf
- 9 Zhang, X., Stamey, J. D., & Mathur, M. B. (2020). Assessing the impact of unmeasured confounders for credible and reliable real-world evidence. *Pharmacoepidemiology and Drug Safety*, 29(10), 1219-1227. <https://doi.org/10.1002/pds.5117>
- 10 Wasserstein, R.L. & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*. 70 (2):. 129-133. doi:10.1080/00031305.2016.1154108
- 11 Source [EMA: Statistical Principles for Clinical Trials], retrieved on <date: 12/14/2022, from https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf
- 12 Abelson, R. P. (1995). *Statistics as principled argument* (First). Psychology Press.
- 13 Newhouse, J. P., & McClellan, M. (1998). ECONOMETRICS IN OUTCOMES RESEARCH: The Use of Instrumental Variables. *Annual Review of Public Health*, 19(1), 17-34. <https://doi.org/10.1146/annurev.publhealth.19.1.17>
- 14 VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, 167(4), 268. <https://www.acpjournals.org/doi/10.7326/M16-2607>
- 15 Desai, R. J., Patorno, E., Vaduganathan, M., Mahesri, M., Chin, K., Levin, R., Solomon, S. D., & Schneeweiss, S. (2021). Effectiveness of angiotensin-NEPRILYSIN inhibitor treatment versus renin-angiotensin system blockade in older adults with heart failure in clinical care. *Heart*, 107(17), 1407-1416. <https://doi.org/10.1136/heartjnl-2021-319405>
- 16 Zhang, X., Faries, D. E., Li, H., Stamey, J. D., & Imbens, G. W. (2018). Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiology and Drug Safety*, 27(4), 373-382. <https://doi.org/10.1002/pds.4394>
- 17 Schneeweiss, S., Eddings, W., Glynn, R. J., Patorno, E., Rassen, J., & Franklin, J. M. (2017). Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. *Epidemiology*, 28(2), 237-248. <https://doi.org/10.1097/ede.0000000000000581>
- 18 Tannen, R. L., Weiner, M. G., & Xie, D. (2009). Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: Comparison of database and randomised controlled trial findings. *BMJ*, 338(jan27 1). <https://doi.org/10.1136/bmj.b81>
- 19 Liu, M., Bunn, V., Hupf, B., Lin, J., & Lin, J. (2021). Propensity-score-based meta-analytic predictive prior for incorporating real-world and Historical Data. *Statistics in Medicine*, 40(22), 4794-4808. <https://doi.org/10.1002/sim.9095>
- 20 Hupf, B., Bunn, V., Lin, J., & Dong, C. (2021). Bayesian semiparametric meta-analytic-predictive prior for historical control borrowing in clinical trials. *Statistics in Medicine*, 40(14), 3385-3399. <https://doi.org/10.1002/sim.8970>

- 21 Ye, T., Ertefaie, A., Flory, J., Hennessy, S., & Small, D. S. (2022). Instrumented difference-in-differences. *Biometrics*. <https://doi.org/10.1111/biom.13783>
- 22 PDUFA VII: Fiscal Years 2023-2027. U.S. Food And Drug Administration. <https://www.fda.gov/industry/prescription-drug-user-fee-amendments/pdufa-vii-fiscal-years-2023-2027>
- 23 McGrath LJ, Spangler L, Curtis JR, et al. Using negative control outcomes to assess the comparability of treatment groups among women with osteoporosis in the United States. *Pharmacoepidemiol Drug Saf*. 2020;29(8):854-863. doi:10.1002/pds.5037
- 24 Spangler L, Nielson C, Brookhart M, Hernandez R, Stad R, Lin J. Myocardial Infarction and Stroke Risks Among Patients Who Initiated Treatment with Denosumab or Zoledronic Acid for Osteoporosis [abstract]. *Arthritis Rheumatol*. 2022; 74 (suppl 9). <https://acrabstracts.org/abstract/myocardial-infarction-and-stroke-risks-among-patients-who-initiated-treatment-with-denosumab-or-zoledronic-acid-for-osteoporosis/> Accessed December 16, 2022
- 25 Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D., & Suchard, M. A. (2018). Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences*, 115(11), 2571-2577. <https://doi.org/10.1073/pnas.1708282114>
- 26 Incerti, D., Bretscher, M. T., Lin, R., & Harbron, C. (2022). A Meta-analytic framework to adjust for bias in external control studies. *Pharmaceutical Statistics*. <https://doi.org/10.1073/pnas.1708282114>
- 27 88 FR 2933 (2023). Understanding the Use of Negative Controls To Assess the Validity of Non-Interventional Studies of Treatment Using Real-World Evidence; Public Workshop. <https://www.federalregister.gov/documents/2023/01/18/2023-00840/understanding-the-use-of-negative-controls-to-assess-the-validity-of-non-interventional-studies-of>